Multi-Player Bandits Revisited

Lilian Besson, Emilie Kaufmann

Multi-Player Bandits Revisited

Lilian Besson[†] LILIAN.BESSON@CENTRALESUPELEC.FR
CentraleSupélec (campus of Rennes), IETR, SCEE Team,
Avenue de la Boulaie – CS 47601*, F-*35576 *Cesson-Sévigné, France*

Emilie Kaufmann EMILIE.KAUFMANN@UNIV-LILLE1.FR
CNRS & Université de Lille, Inria SequeL team
*UMR 9189 – CRIStAL, F-*59000 *Lille, France*

Abstract

Multi-player Multi-Armed Bandits (MAB) have been extensively studied in the literature, motivated by applications to Cognitive Radio systems. Driven by such applications as well, we motivate the introduction of several levels of feedback for multi-player MAB algorithms. Most existing work assume that *sensing information* is available to the algorithm. Under this assumption, we improve the state-of-the-art lower bound for the regret of any decentralized algorithms and introduce two algorithms, $\mathrm{RandTopM}$ and MCTopM, that are shown to empirically outperform existing algorithms. Moreover, we provide strong theoretical guarantees for these algorithms, including a notion of asymptotic optimality in terms of the number of selections of bad arms. We then introduce a promising heuristic, called $\mathrm{Selfish}$, that can operate without sensing information, which is crucial for emerging applications to Internet of Things networks. We investigate the empirical performance of this algorithm and provide some first theoretical elements for the understanding of its behavior.

Keywords: Multi-Armed Bandits; Decentralized algorithms; Reinforcement learning; Cognitive Radio; Opportunistic Spectrum Access.

1. Introduction

Several sequential decision making problems under the constraint of partial information have been studied since the 1950s under the name of Multi-Armed Bandit (MAB) problems (Robbins, 1952; Lai and Robbins, 1985). In a stochastic MAB model, an agent is facing K unknown probability distributions, called arms in reference to the arms of a one-armed bandit (or slot machine) in a casino. Each time she selects (or draws) an arm, she receives a reward drawn from the associated distribution. Her goal is to build a sequential selection strategy that maximizes the total reward received. A class of algorithms to solve this problem is based on Upper Confidence Bounds (UCB), first proposed by Lai and Robbins (1985); Agrawal (1995) and further popularized by Auer et al. (2002a). The field has been very active since then, with several algorithms proposed and analyzed, both theoretically and empirically, even beyond the stochastic assumption on arms, as explained in the survey by Bubeck et al. (2012).

The initial motivation to study MAB problems arose from clinical trials (the first MAB model can be traced back to 1933, by Thompson), in which a doctor sequentially allocates treatments (arms) to patients and observes their efficacy (reward). More recently, applications of MAB have shifted towards sequential content recommendation, *e.g.* sequential display of advertising to customers or A/B testing (Li et al., 2010; Chapelle et al., 2014). In the mean time, MAB were found to be relevant to the field of Cognitive Radio (CR, Mitola and Maguire (1999)), and Jouini et al. (2009, 2010) first proposed to use UCB_1 for the Opportunistic Spectrum Access (OSA) problem, and successfully conducted experiments on real radio

networks demonstrating its usefulness. For CR applications, each arm models the quality or availability of a radio channel (a frequency band) in which there is some background traffic (*e.g.*, primary users paying to have a guaranteed access to the channel in the case of OSA). A smart radio device needs to insert itself in the background traffic, by sequentially choosing a channel to access and try to communicate on, seeking to optimize the quality of its global transmissions.

For the development of CR, a crucial step is to insert *multiple* $M \geq 2$ smart devices in the *same* background traffic. With the presence of a central controller that can assign the devices to separate channels, this amounts to choosing at each time step *several* arms of a MAB in order to maximize the global rewards, and can thus be viewed as an application of the multiple-play bandit, introduced by Anantharam et al. (1987) and recently studied by Komiyama et al. (2015). Due to the communication cost implied by a central controller, a more relevant model is the *decentralized multi-player* multi-armed bandit model, introduced by Liu and Zhao (2010) and Anandkumar et al. (2010, 2011), in which players select arms individually and collisions may occur, that yield a loss of reward. Further algorithms were proposed in similar models by Tekin and Liu (2012) and Kalathil et al. (2012) (under the assumption that each arm is a Markov chain) and by Avner and Mannor (2015, 2016) and Rosenski et al. (2016) (for *i.i.d.* or piece-wise *i.i.d.* arms). The goal for every player is to select most of the time one of the M best arms, without colliding too often with other players. A first difficulty relies in the well-known trade-off between *exploration* and *exploitation*: players need to explore all arms to estimate their means while trying to focus on the best arms to gain as much rewards as possible. The decentralized setting considers no exchange of information between players, that only know K and M, and to avoid collisions, players should furthermore find orthogonal configurations (*i.e.*, the M players use the M best arms without any collision), without communicating. Hence, in that case the trade-off is to be found between exploration, exploitation *and* low collisions.

All these above-mentioned works are motivated by the OSA problem, in which it is assumed that *sensing* occurs, that is each smart device observes the availability of a channel (sample from the arm) *before* trying to transmit and possibly experiment a collision with other smart devices. However some real radio networks do not use sensing at all, *e.g.*, emerging standards developed for *Internet of Things* (IoT) networks such as LoRaWAN. Thus, to take into account these new applications, algorithms with additional constraints on the available feedback have to be proposed within the multiple-player MAB model. Especially, the typical approach that combines a (single-player) bandit algorithm based on the sensing information –to learn the quality of the channels while targeting the best ones– with a low-complexity decentralized collision avoidance protocol, is no longer possible.

In this paper, we take a step back and present the different feedback levels possible for multi-player MAB algorithms. For each of them, we propose algorithmic solutions supported by both experimental and theoretical guarantees. In the presence of sensing information, our contributions are a new problem-dependent regret lower bound, tighter than previous work, and the introduction of two algorithms, $\mathrm{RandTopM}$ and MCTopM. Both are shown to achieve an asymptotically optimal number of selections of the sub-optimal arms, and for MCTopM we furthermore establish a logarithmic upper bound on the regret, that follows from a careful control of the number of collisions. In the absence of sensing information, we propose the $\mathrm{Selfish}$ heuristic and investigate its performance. Our study of this algorithm is supported by (promising) empirical performance and some first (disappointing) theoretical elements.

The rest of the article is organized as follows. We introduce the multi-player bandit model with three feedback levels in Section 2, and give a new regret lower bound in Section 3. The $\mathrm{RandTopM}$, MCTopM and $\mathrm{Selfish}$ algorithms are introduced in Section 4, with the result of our experimental study reported in Section 5. Theoretical elements are then presented in Section 6.

2. Multi-Player Bandit Model with Different Feedback Levels

We consider a K-armed Bernoulli bandit model, in which arm k is a Bernoulli distribution with mean $\mu_k \in [0,1]$. We denote $(Y_{k,t})_{t\in\mathbb{N}}$ the *i.i.d.* (binary) *reward stream* for arm k, that satisfies $\mathbb{P}(Y_{k,t}=1)=\mu_k$ and that is independent from the other rewards streams. However we mention that our lower bound and all our algorithms (and their analysis) can be easily extended to one-dimensional exponential families (just like for the kl-UCB algorithm of Cappé et al. (2013)). For simplicity, we focus on the Bernoulli case, that is also the most relevant for Cognitive Radio, as it can model channel availabilities.

In the multi-player MAB setting, there are $M \in \{1,\dots,K\}$ players (or agents), that have to make decisions at some pre-specified time instants. At time step $t \in \mathbb{N}, t \geq 1$, player j selects an arm $A^j(t)$, independently from the other players' selections. A *collision* occurs at time t if at least two players choose the same arm. We introduce the two events, for $j \in \{1,\dots,M\}$ and $k \in \{1,\dots,K\}$,

$$C^j(t) := \{\exists j' \neq j : A^{j'}(t) = A^j(t)\} \quad \text{and} \quad C_k(t) := \left\{\#\{j : A^j(t)=k\} > 1\right\}, \tag{1}$$

that respectively indicate that a collision occurs at time t for player j and that a collision occurs at time t on arm k. Each player j then receives (and observes) the *binary rewards* $r^j(t) \in \{0,1\}$,

$$r^j(t) := Y_{A^j(t),t}\, \mathbb{1}(\overline{C^j(t)}). \tag{2}$$

In words, she receives the reward of the selected arm if she is the only one to select this arm, and a reward zero otherwise[1]. Other models for rewards loss have been proposed in the literature (*e.g.*, the reward is randomly allocated to one of the players selecting it), but we focus on full reward occlusion in this article.

A multi-player MAB strategy is a tuple $\rho = (\rho^1,\dots,\rho^M)$ of arm selection strategies for each player, and the goal is to propose a strategy that maximizes the total reward of the system, under some constraints. First, each player j should adopt a *sequential* strategy ρ^j, that decides which arm to select at time t based on *previous observations*. Previous observations for player j at time t always include the previously chosen arms $A^j(s)$ and received rewards $r^j(s)$ for $s<t$, but may also include the *sensing information* $Y_{A^j(t),t}$ or the *collision information* $C^j(t)$. More precisely, depending on the application, one may consider the following three observation models, (I), (II) and (III).

(I) **Simultaneous sensing and collision**: player j observes $Y_{A^j(t),t}$ *and* $C^j(t)$ (not previously studied).

(II) **Sensing, then collision**: player j observes $Y_{A^j(t),t}$, *then* observes the reward, and thus also $C^j(t)$ only if $Y_{A^j(t),t}=1$. This common setup, studied for example by Anandkumar et al. (2011); Avner and Mannor (2015); Rosenski et al. (2016), is relevant to model the OSA problem: the device first checks for the presence of primary users in the chosen channel. If this channel is free ($Y_{A^j(t),t}=1$), the transmission is successful ($r^j(t)=1$) if no collision occurs with other smart devices ($\overline{C^j(t)}$).

(III) **No sensing**: player j only observes the reward $r^j(t)$. For IoT networks, this reward can be interpreted as an acknowledgement from a Base Station, received when a communication was successful. A lack of acknowledgment may be due to a collision with a device from the background traffic ($Y_{A^j(t),t}=0$), or to a collision with one of the others players ($C^j(t)$). However, the sensing and collision information are censored. Recently, Bonnefoi et al. (2017) presented the first (bandit-based) algorithmic solutions under this (harder) feedback model, in a slightly different setup, more suited to large scale IoT applications.

1. This provides another reason to focus on the Bernoulli model. It is the hardest model, in the sense that receiving a reward zero is not enough to detect collisions. For other models, the data streams $(Y_{k,s})_s$ are usually continuously distributed, with no mass at zero. Hence receiving $r^j(t)=0$ directly gives $\mathbb{1}(C^j(t))=1$.

Under each of these three models, we define $\mathcal{F}_t^j$ to be the filtration generated by the observations gathered by player j up to time t (which contains different information under models (I), (II) and (III)). While a *centralized* algorithm may select the vector of actions for all players $(A^1(t), \dots, A^M(t))$ based on all the observations from $\bigcup_j \mathcal{F}_{t-1}^j$, under a *decentralized* algorithm the arm selected at time t by player j only depends on the past observation of this player. More formally, $A^j(t)$ is assumed to be $\mathcal{F}_{t-1}^j$-measurable.

Definition 1. *We denote by μ_1^* the best mean, μ_2^* the second best etc, and by M-best the (non-sorted) set of the indices of the M arms with largest mean (*best arms*): if $\mu_1^* = \mu_{k_1}, \dots, \mu_M^* = \mu_{k_M}$ then M-best $= \{k_1, \dots, k_M\}$. Similarly, M-worst denotes the set of indices of the $K - M$ arms with smallest means (*worst arms*), $\{1, \dots, K\} \setminus M$-best. Note that they are both uniquely defined if $\mu_M^* > \mu_{M+1}^*$.*

Following a natural approach in the bandit literature, we evaluate the performance of a multi-player strategy using the *expected regret* (later simply referred to as regret), that measures the performance gap with respect to the best possible strategy. The regret of the strategy ρ at horizon T is the difference between the cumulated reward of an oracle strategy, assigning in this case the M players to M-best, and the cumulated reward of strategy ρ:

$$R_T(\boldsymbol{\mu}, M, \rho) := \left(\sum_{k=1}^{M} \mu_k^*\right) T - \mathbb{E}_\mu\left[\sum_{t=1}^{T}\sum_{j=1}^{M} r^j(t)\right]. \tag{3}$$

Maximizing the expected sum of global reward of the system is indeed equivalent to minimizing the regret, and we now investigate the best possible *regret rate* of a decentralized multi-player algorithm.

3. An Asymptotic Regret Lower Bound

In this section, we provide a useful decomposition of the regret (Lemma 3) that permits to establish a new problem-dependent lower bound on the regret (Theorem 6), and also provides key insights on the derivation of regret upper bounds (Lemma 7).

3.1 A Useful Regret Decomposition

We introduce additional notations in the following definition.

Definition 2. *Let $T_k^j(T) := \sum_{t=1}^{T} \mathbb{1}(A^j(t) = k)$, and denote $T_k(T) := \sum_{j=1}^{M} T_k^j(T)$ the* number of selections *of arm $k \in \{1, \dots, K\}$ by any player $j \in \{1, \dots, M\}$, up to time T.*

Let $C_k(T)$ be the number of colliding players[2] *on arm $k \in \{1, \dots, K\}$ up to horizon T:*

$$C_k(T) := \sum_{t=1}^{T}\sum_{j=1}^{M} \mathbb{1}(C^j(t))\mathbb{1}(A^j(t) = k). \tag{4}$$

Letting $\mathcal{P}_M = \left\{\boldsymbol{\mu} \in [0,1]^K : \mu_M^* > \mu_{M+1}^*\right\}$ be the set of bandit instances such that there is a strict gap between the M best arms and the other arms, we now provide a regret decomposition for any $\boldsymbol{\mu} \in \mathcal{P}_M$.

Lemma 3. *For any bandit instance $\boldsymbol{\mu} \in \mathcal{P}_M$ such that $\mu_M^* > \mu_{M+1}^*$, it holds that* *(Proved in App. A.1)*

$$R_T(\boldsymbol{\mu}, M, \rho) = \underbrace{\sum_{k \in M\text{-worst}} (\mu_M^* - \mu_k)\mathbb{E}_\mu[T_k(T)]}_{\text{(a)}} + \underbrace{\sum_{k \in M\text{-best}} (\mu_k - \mu_M^*)(T - \mathbb{E}_\mu[T_k(T)])}_{\text{(b)}} + \underbrace{\sum_{k=1}^{K} \mu_k \mathbb{E}_\mu[C_k(T)]}_{\text{(c)}}.$$

2. When n players choose arm k at time t, this counts as n collisions, not just one. So $C_k(T)$ counts the total *number of colliding players* rather than the number of collision events. Hence there is small abuse of notation when calling it a number of collisions.

In this decomposition, term (a) counts the lost rewards due to *sub-optimal arms* selections ($k \in M$-worst), term (b) counts the number of times the *best arms* were *not* selected ($k \in M$-best), and term (c) counts the weighted number of collisions, on *all arms*. It is valid for both centralized and decentralized algorithms. For centralized algorithms, due to the absence of collisions, (c) is obviously zero, and (b) is non-negative, as $T_k(T) \leq T$. For decentralized algorithms, (c) may be significantly large, and term (b) may be negative, as many collisions on arm k may lead to $T_k(T) > T$. However, a careful manipulation of this decomposition (see Appendix A.2) shows that the regret is always lower bounded by term (a).

Lemma 4. *For any strategy ρ and $\boldsymbol{\mu} \in \mathcal{P}_M$, it holds that* $R_T(\boldsymbol{\mu}, M, \rho) \geq \sum_{k \in M\text{-worst}} (\mu_M^\star - \mu_k)\mathbb{E}_\mu[T_k(T)]$.

3.2 An Improved Asymptotic Lower Bound on the Regret

To express our lower bound, we need to introduce $\mathrm{kl}(x, y)$ as the Kullback-Leibler divergence between the Bernoulli distribution of mean $x \neq 0, 1$ and that of mean $y \neq 0, 1$, so that $\mathrm{kl}(x, y) := x \log(x/y) + (1 - x) \log((1 - x)/(1 - y))$. We first introduce the assumption under which we derive a regret lower bound, that generalizes a classical assumption made by Lai and Robbins (1985) in single-player bandit models.

Definition 5. *A strategy ρ is* **strongly uniformly efficient** *if for all $\boldsymbol{\mu} \in \mathcal{P}_M$ and for all $\alpha \in (0, 1)$,*

$$R_T(\boldsymbol{\mu}, M, \rho) \underset{T \to +\infty}{=} o(T^\alpha) \quad \textit{and} \quad \forall j \in \{1, \dots, M\}, k \in M\text{-best}, \ \frac{T}{M} - \mathbb{E}_\mu[T_k^j(T)] \underset{T \to +\infty}{=} o(T^\alpha). \quad (5)$$

Having a small regret on every problem instance, *i.e.*, uniform efficiency, is a natural assumption for algorithms, that rules out algorithms tuned to perform well on specific instances only. From this assumption ($R_T(\boldsymbol{\mu}, M, \rho) = o(T^\alpha)$) and the decomposition of Lemma 3 one can see[3] that for every $k \in M$-best, $T - \mathbb{E}_\mu[T_k(T)] = o(T^\alpha)$, and so

$$\sum_{j=1}^{M} \left(\frac{T}{M} - \mathbb{E}_\mu[T_k^j(T)] \right) = o(T^\alpha). \quad (6)$$

The additional assumption in (5) further implies some notion of *fairness*, as it suggests that each of the M players spends on average the same amount of time on each of the M best arms. Note that this assumption is satisfied by any strategy that is invariant under every permutation of the players, *i.e.*, for which the distribution of the observations under $\rho^\gamma = (\rho^{\gamma(1)}, \dots, \rho^{\gamma(M)})$ is independent from the choice of permutation $\gamma \in \Sigma_M$. In that case, it holds that $\mathbb{E}_\mu[T_k^j(T)] = \mathbb{E}_\mu[T_k^{j'}(T)]$ for every arm k and $(j, j') \in \{1, \dots, M\}$, hence (5) and (6) are equivalent, and strong uniform efficiency is equivalent to standard uniform efficiency. Note that all our proposed algorithms are permutation invariant and MCTopM is thus an example of strongly uniformly efficient algorithm, as we prove in Section 6 that its regret is logarithmic on every instance $\mu \in \mathcal{P}_M$.

We now state a problem-dependent asymptotic lower bound on the number of sub-optimal arms selections under a *decentralized strategy* that has access to the sensing information. This result, proved in Appendix B, yields an asymptotic logarithmic lower bound on the regret, also given in Theorem 6.

Theorem 6. *Under observation models (I) and (II), for any strongly uniformly efficient* decentralized *policy ρ and $\boldsymbol{\mu} \in \mathcal{P}_M$,*

$$\forall j \in \{1, \dots, M\}, \ \forall k \in M\text{-worst}, \quad \liminf_{T \to \infty} \frac{\mathbb{E}_\mu[T_k^j(T)]}{\log(T)} \geq \frac{1}{\mathrm{kl}(\mu_k, \mu_M^\star)}. \quad (7)$$

3. With some arguments used in the proof of Lemma 4 to circumvent the fact that (b) may be negative.

From Lemma 4, it follows that

$$\liminf_{T\to+\infty} \frac{R_T(\boldsymbol{\mu}, M, \rho)}{\log(T)} \geq M \times \left(\sum_{k \in M\text{-worst}} \frac{(\mu_M^* - \mu_k)}{\mathrm{kl}(\mu_k, \mu_M^*)} \right). \tag{8}$$

Observe that the regret lower bound (8) is tighter than the state-of-the-art lower bound in this setup given by Liu and Zhao (2010), that states that

$$\liminf_{T\to+\infty} \frac{R_T(\boldsymbol{\mu}, M, \rho)}{\log(T)} \geq \sum_{k \in M\text{-worst}} \left(\sum_{j=1}^{M} \frac{(\mu_M^* - \mu_k)}{\mathrm{kl}(\mu_k, \mu_j^*)} \right), \tag{9}$$

as for every $k \in M$-worst and $j \in \{1, \dots, M\}$, $\mathrm{kl}(\mu_k, \mu_j^*) \geq \mathrm{kl}(\mu_k, \mu_M^*)$ (see Figure 5 in Appendix F.1). It is worth mentioning that Liu and Zhao (2010) proved a lower bound under the more general assumption for ρ that there exists some numbers (a_k^j) such that $a_k^j T - \mathbb{E}_\mu[T_k^j(T)] = o(T^\alpha)$ whereas in Definition 5 we make the choice $a_k^j = 1/M$. Our result could be extended to this case but we chose to keep the notation simple and focus on *fair allocation* of the optimal arms between players.

Interestingly, our lower bound is exactly a multiplicative constant factor M away from the lower bound given by Anantharam et al. (1987) for centralized algorithms (which is clearly a simpler setting). This intuitively suggests the number of players M as the (multiplicative) *"price of decentralized learning"*. However, to establish our regret bound, we lower bounded the number of collisions by zero, which may be too optimistic. Indeed, for an algorithm to attain the lower bound (8), the number of selections of each sub-optimal arm should match the lower bound (7) *and* term (b) and term (c) in the regret decomposition of Lemma 3 should be negligible compared to $\log(T)$. To the best of our knowledge, no algorithm has been shown to experience only $o(\log(T))$ collisions so far, for every $M \in \{2, \dots, K\}$ and $\boldsymbol{\mu} \in \mathcal{P}_M$.

A lower bound on the minimal number of collisions experienced by any strongly uniformly efficient decentralized algorithm would thus be a nice complement to our Theorem 6, and it is left as future work.

3.3 Towards Regret Upper Bounds

A natural approach to obtain an upper bound on the regret of an algorithm is to upper bound separately each of the three terms defined in Lemma 3. The following result shows that term (b) can be related to the number of sub-optimal selections and the number of collisions that occurs on the M best arms.

Lemma 7. *The term (b) in Lemma 3 is upper bounded as* *(Proved in Appendix A.3)*

$$(b) \leq (\mu_1^* - \mu_M^*) \Big(\sum_{k \in M\text{-worst}} \mathbb{E}_\mu[T_k(T)] + \sum_{k \in M\text{-best}} \mathbb{E}_\mu[C_k(T)] \Big). \tag{10}$$

This result can also be used to recover Proposition 1 from Anandkumar et al. (2011), giving an upper bound on the regret that only depends on the *expected number of sub-optimal selections* – $\mathbb{E}_\mu[T_k(T)]$ for $k \in M$-worst – and the *expected number of colliding players on the optimal arms* – $\mathbb{E}_\mu[C_k(T)]$ for $k \in M$-best. Note that, in term (c) the number of colliding players on the sub-optimal arm k may be upper bounded as $\mathbb{E}_\mu[C_k(T)] \leq M\mathbb{E}_\mu[T_k(T)]$.

In the next Section, we present an algorithm that has a logarithmic regret, while ensuring that the number of sub-optimal selections is matching the lower bound of Theorem 6.

4. New Algorithms for Multi-Player Bandits

When sensing is possible, that is under observation models (I) and (II), most existing strategies build on a *single-player bandit algorithm* (usually an *index policy*) that relies on the sensing information, together with an *orthogonalization strategy* to deal with collisions. We present this approach in more details in Section 4.1 and introduce two new algorithms of this kind, RandTopM and MCTopM. Then, we suggest in Section 4.2 a completely different approach, called Selfish, that no longer requires an orthogonalization strategy as the collisions are directly accounted for in the indices that are used. Selfish can also be used under observation model (III) –*without sensing*–, and without the knowledge of M.

4.1 Two New Strategies Based on Indices and Orthogonalization: RandTopM and MCTopM

In a single-player setting, *index policies* are popular bandit algorithms: at each round one index is computed for each arm, that only depends on the history of plays of this arm and (possibly) some exogenous randomness. Then, the arm with highest index is selected. This class of algorithms includes the UCB family, in which the index of each arm is an Upper Confidence Bound for its mean, but also some Bayesian algorithms like Bayes-UCB (Kaufmann et al., 2012a) or the randomized Thompson Sampling algorithm (Thompson, 1933; Agrawal and Goyal, 2012; Kaufmann et al., 2012b).

The approaches we now describe for multi-player bandits can be used in combination with any index policy, but we restrict our presentation to UCB algorithms, for which strong theoretical guarantees can be obtained. In particular, we focus on two types of indices: UCB_1 indices (Auer et al., 2002a) and kl-UCB indices (Cappé et al., 2013), that can be defined for each player j in the following way. Letting $S_k^j(t) := \sum_{s=1}^t Y_{k,s} \mathbb{1}(A^j(t) = k)$ the current sum of sensing information obtained by player j for arm k, $\widehat{\mu}_k^j(t) = S_k^j(t)/T_k^j(t)$ (if $T_k^j(t) \neq 0$) is the empirical mean of arm k for player j and one can define the index

$$g_k^j(t) := \begin{cases} \widehat{\mu}_k^j(t) + \sqrt{f(t)/(2T_k^j(t))} & \text{for } \mathrm{UCB}_1, \\ \sup\left\{ q \in [0,1] : T_k^j(t) \times \mathrm{kl}(\widehat{\mu}_k^j(t), q) \leq f(t) \right\} & \text{for kl-UCB}, \end{cases} \tag{11}$$

where $f(t)$ is some *exploration function*. $f(t)$ is usually taken to be $\log(t)$ in practice, and slightly larger in theory, which ensures that $\mathbb{P}(g_k^j(t) \geq \mu_k) \gtrsim 1 - 1/t$ (see Cappé et al. (2013)). A classical (single-player) UCB algorithm aims at the arm with largest index. However, if each of the M players selects the arm with largest UCB, all the players will end up colliding most of the time on the best arm. To circumvent this problem, several coordination mechanisms have emerged, that rely on *ordering* the indices and targeting *one of* the M-best indices.

While the TDFS algorithm (Liu and Zhao, 2010) relies on the player agreeing in advance on the time steps at which they will target each of the M best indices (even though some alternative without pre-agreement are proposed), the RhoRand algorithm (Anandkumar et al., 2011) relies on randomly selected *ranks*. More formally, letting $\pi(k, \mathbf{g})$ be the index of the k-th largest entry in a vector $\mathbf{g}$, in RhoRand each player maintains at time t an internal rank $R^j(t) \in \{1, \dots, M\}$ and selects at time t,

$$A^j(t) := \pi\left(R^j(t), [g_\ell^j(t)]_{\ell=1,\dots,K}\right). \tag{12}$$

If a collision occurs, a new rank is drawn uniformly at random: $R^j(t+1) \sim \mathcal{U}(\{1, \dots, M\})$.

We now propose two alternatives to this strategy, that do not rely on ranks and rather randomly fix themselves on one *arm* in $\widehat{M^j}(t)$, that is defined as the set of arms that have the M largest indices:

$$\widehat{M^j}(t) := \left\{ \pi\left(k, \{g_\ell^j(t)\}_{\ell=1,\dots,K}\right), k = 1, \dots, M \right\}. \tag{13}$$

Our proposal MCTopM is stated below as Algorithm 1, while a simpler variant, called RandTopM, is stated as Algorithm 2 in Appendix C. We focus on MCTopM as it is easier to analyze and performs better. Both algorithms ensure that player j always selects at time $t+1$ an arm from $\widehat{M^j}(t)$. When a collision occurs RandTopM randomly switches arm within $\widehat{M^j}$, while MCTopM uses a more sophisticated mechanism, that is reminiscent of "Musical Chair" (MC) and inspired by the work of Rosenski et al. (2016): players tend to fix themselves on arms ("chairs") and ignore future collision when this happens.

1 Let $A^j(1) \sim \mathcal{U}(\{1,\dots,K\})$ and $C^j(1) = \text{False}$ and $s^j(1) = \text{False}$
2 **for** $t = 0,\dots,T-1$ **do**
3 **if** $A^j(t) \notin \widehat{M^j}(t)$ **then** // transition (3) or (5)
4 $A^j(t+1) \sim \mathcal{U}\left(\widehat{M^j}(t) \cap \left\{k : g_k^j(t-1) \le g_{A^j(t)}^j(t-1)\right\}\right)$ // not empty
5 $s^j(t+1) = \text{False}$ // aim at an arm with a smaller UCB at $t-1$
6 **else if** $C^j(t)$ and $\overline{s^j(t)}$ **then** // collision and not fixed
7 $A^j(t+1) \sim \mathcal{U}\left(\widehat{M^j}(t)\right)$ // transition (2)
8 $s^j(t+1) = \text{False}$
9 **else** // transition (1) or (4)
10 $A^j(t+1) = A^j(t)$ // stay on the previous arm
11 $s^j(t+1) = \text{True}$ // become or stay fixed on a ``chair''
12 **end**
13 Play arm $A^j(t+1)$, get new observations (sensing and collision),
14 Compute the indices $g_k^j(t+1)$ and set $\widehat{M^j}(t+1)$ for next step.
15 **end**

Algorithm 1: The MCTopM decentralized learning policy (for a fixed underlying index policy g^j).

More precisely, under MCTopM, if player j did not encounter a collision when using arm k at time t, then she marks her current arm as a "chair" ($s^j(t+1) = \text{True}$), and will keep using it even if collisions happen in the future (Lines 9-11). As soon as this "chair" k is no longer in $\widehat{M_j}(t)$, a new arm is sampled uniformly from a subset of $\widehat{M^j}(t)$, defined with the previous indices $g^j(t-1)$ (Lines 3-5). The subset enforces a certain inequality on indices, $g_{k'}^j(t-1) \le g_k^j(t-1)$ and $g_{k'}^j(t) \ge g_k^j(t)$, when switching from $k = A^j(t)$ to $k' = A^j(t+1)$. This helps to control the number of such changes of arm, as shown in Lemma 9. The considered subset is never empty as it contains at least the arm replacing the $k \in \widehat{M^j}(t-1)$ in $\widehat{M^j}(t)$. Collisions are dealt with only for non-fixed player j, and when the previous arm is still in $\widehat{M^j}(t)$. In this case, a new arm is sampled uniformly from $\widehat{M^j}(t)$ (Lines 6-8). This stationary aspect helps to minimize the number of collisions, as well as the number of switches of arm. The five different transitions (1), (2), (3), (4), (5) refer to the notations used in the analysis of MCTopM (see Figure 3 in Appendix D.3).

4.2 The Selfish Approach

Under observation model (III) no sensing information is available and the previous algorithms cannot be used, as the sum of sensing information $S_k^j(t)$ and thus the empirical mean $\widehat{\mu}_k^j(t)$ cannot be computed, hence neither the indices $g_k^j(t)$. However, one can still define a notion of *empirical reward* received from arm k by player j, by introducing

$$\widetilde{S_k}^j(t) = \sum_{t=1}^{T} r^j(t) \mathbb{1}(A^j(t) = k) \quad \text{and letting} \quad \widetilde{\mu_k}^j(t) := \widetilde{S_k}^j(t) \,/\, T_k^j(t). \tag{14}$$

Note that $\widetilde{\mu_k}^j(t)$ is no longer meant to be an unbiased estimate of μ_k as it also takes into account the collision information, that is present in the reward. Based on this empirical reward, one can similarly defined modified indices as

$$\widetilde{g_k}^j(t) = \begin{cases} \widetilde{\mu_k}^j(t) + \sqrt{f(t)/(2T_k^j(t))} & \text{for } \mathrm{UCB}_1, \\ \sup\left\{q \in [0,1] : T_k^j(t) \times \mathrm{kl}(\widetilde{\mu_k}^j(t), q) \leq f(t)\right\} & \text{for kl-UCB.} \end{cases} \tag{15}$$

Given any of these two index policies (UCB_1 or kl-UCB), the Selfish algorithm is defined by,

$$A^j(t) = \underset{k\in\{1,\dots,K\}}{\operatorname{argmax}} \ \widetilde{g_k}^j(t-1). \tag{16}$$

The name comes from the fact that each player is targeting, in a "selfish" way, the arm that has the highest index, instead of accepting to target only one of the M best. The reason that this may work precisely comes from the fact that $\widetilde{g_k}^j(t)$ is no longer an upper-confidence on μ_k, but some hybrid index that simultaneously increases when a transmission occurs and decreases when a collision occurs.

This behavior is easier to be understood for the case of Selfish-UCB_1 in which, letting $N_k^{j,C}(t) = \sum_{s=1}^t \mathbb{1}(C^j(t))$ be the number of collisions on arm k, one can show that the hybrid Selfish index induces a penalty proportional to the fraction of collision on this arm and the quality of the arm itself:

$$\widetilde{g_k}^j(t) = g_k^j(t) - \underbrace{\left(\frac{N_k^{j,C}(t)}{N_k^j(t)}\right)}_{\text{fraction of collisions}} \underbrace{\left(\frac{1}{N_k^{j,C}(t)} \sum_{t=1}^{T} Y_{A^j(t),t} \mathbb{1}(C^j(t)) \mathbb{1}(A^j(t) = k)\right)}_{\text{estimate of } \mu_k}. \tag{17}$$

From a bandit perspective, it looks like each player is using a stochastic bandit algorithm (UCB_1 or kl-UCB) when interacting with K arms that give a feedback (the reward, and not the sensing information) that is far from being *i.i.d.* from some distribution, due to the collisions. As such, the algorithm does not appear to be well justified, and one may rather want to use adversarial bandit algorithms like EXP3 (Auer et al., 2002b), that do not require a stochastic (*i.i.d.*) assumption on arms. However, we found out empirically that Selfish is doing surprisingly well, as already noted by Bonnefoi et al. (2017), who did some experiments in the context of IoT applications. We show in Section 6 that Selfish does have a (very) small probability to fail (badly), for some problem with small K, which precludes the possibility of a logarithmic regret for any problem. However, in most cases it empirically performs similarly to all the algorithms described before, and usually outperforms RhoRand, even if it neither exploits the sensing information, nor the knowledge of the number of players M. As such, practitioners may still be interested by the algorithm, especially for Cognitive Radio applications in which sensing is hard or not considered.

5. Empirical performances

We illustrate here the empirical performances of the algorithms presented in Section 4, used in combination with the kl-UCB indices. Some plots are at pages 10 and 11 and most of them in Appendix F.2.

In experiments that are not reported here, we could observe that using kl-UCB rather than UCB_1 indices always yield better practical performance. As the prupose of this work is not to optimize on the index policy, but rather propose new ways of using indices in a decentralized setting, we only report results for kl-UCB. In a first set of experiments, MCTopM, RandTopM and Selfish are benchmarked against the state-of-the-art RhoRand algorithm. We also include a centralized multiple-play kl-UCB algorithm, essentially to check that the *"price of decentralized learning"* is not too large.

We present results for two bandit instance: one with $K = 3$ arms and means $\boldsymbol{\mu} = [0.1, 0.5, 0.9]$, for which two cases $M = 2$ and $M = 3$ are presented in Figure 8. For the second instance $K = 9$ and $\boldsymbol{\mu} = [0.1, 0.2, \dots, 0.9]$, three cases are presented: $M = 6$ in Figure 2, and for the two limit cases $M = 2$ and $M = 9 = K$ in Figure 13. Performance is measured with the *expected* regret up to horizon $T = 10000$, estimated based on 1000 repetitions on the same bandit instance. We also include histograms showing the *distribution* of regret at $t = T$, as this allows to check if the regret is indeed small for *each* run of the simulation. For the plots showing the regret, our *asymptotic* lower bound from Theorem 6 is displayed.

Experiments with a different problem for each repetition (uniformly sampled $\boldsymbol{\mu} \sim \mathcal{U}([0,1]^K)$), are also considered, in Figure 1 and 11. This helps to check that no matter the *complexity* of the considered problem (one measure of complexity being the constant in our lower bound), MCTopM performs similarly or better than all the other algorithms, and Selfish outperforms RhoRand in most cases. Empirically, our proposals were found to almost always outperform RhoRand, and except for Selfish that can fail badly on problems with small K, we verified that MCTopM outperforms the state-of-the-art algorithms in many different problems, and is more and more efficient as M and K grows.

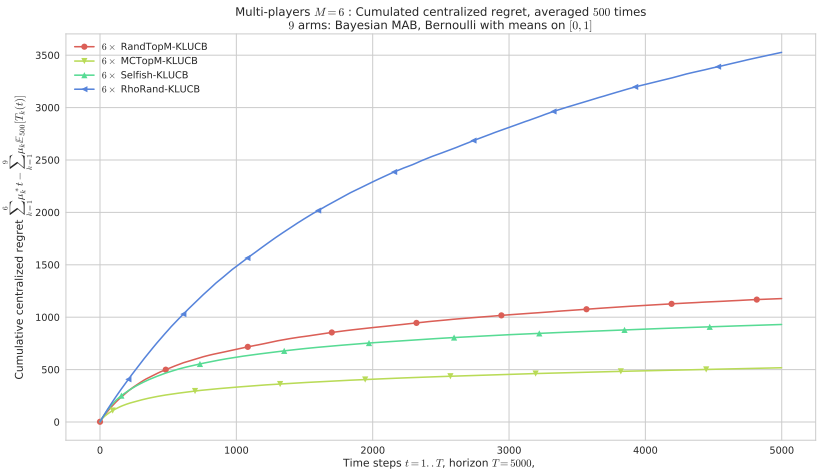


Figure 1: Regret, $M = 6$ players, $K = 9$ arms, horizon $T = 5000$, against 500 problems $\boldsymbol{\mu}$ uniformly sampled in $[0,1]^K$. RhoRand (top blue curve) is outperformed by the other algorithms (and the gain increases with M). MCTopM (bottom yellow) outperforms all the other algorithms is most cases.

In the presence of sensing (observation model (II)), we also compared our algorithms to with MEGA (Avner and Mannor, 2015) and Musical Chair (Rosenski et al., 2016). Yet these two algorithms were found hard to use efficiently in practice and we show in Figure 7 that they perform poorly in comparison to RhoRand, RandTopM and MCTopM. MEGA needs a careful tuning of *five* parameters (c, d, p_0, α and β) to attain reasonable performances. No good guideline for tuning them is provided and using cross validation, as suggested, can be considered out of the scope of online sequential learning. Musical Chair consists of an random exploration phase of length T_0 after which the players quickly converge to orthogonal strategies targeting the M best arms. With probability $1 - \delta$, its regret is proved to be "constant" (of order $\log(1/\delta)$). The theoretical minimal value for T_0 depends on δ, on the horizon T and on a lower bound ϵ on the gap $\Delta = \mu^*_M - \mu^*_{M+1}$, and the practical tuning is hard too.

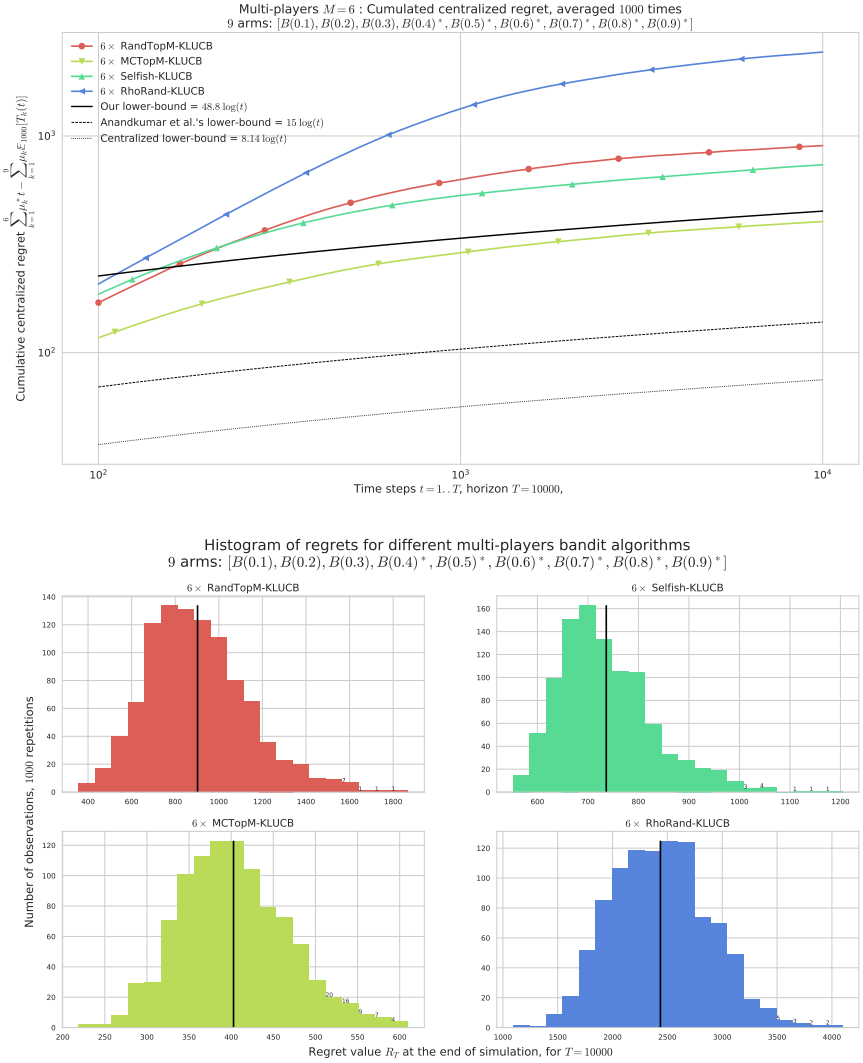


Figure 2: Regret (in $\log\log$ scale), for $M = 6$ players for $K = 9$ arms, horizon $T = 5000$, for 1000 repetitions on problem $\boldsymbol{\mu} = [0.1, \dots, 0.9]$. RandTopM (yellow curve) outperforms Selfish (green), both clearly outperform RhoRand. The regret of MCTopM is logarithmic, empirically with the same slope as the lower bound. The x axis on the regret histograms have different scale for each algorithm.

6. Theoretical elements

Section 6.1 gives an asymptotically optimal analysis of the expected number of sub-optimal draws for RandTopM, MCTopM and RhoRand combined with kl-UCB indices, and Section 6.2 proves that the number of collisions, hence the regret of MCTopM are also logarithmic. Section 6.3 shortly discusses a disappointing result regarding Selfish, with more insights provided in Appendix E.

6.1 Common Analysis for RandTopM- and MCTopM-kl-UCB

Lemma 8 gives a finite-time upper bound on the expected number of draws of a sub-optimal arm k for any player j, that holds for both RandTopM-kl-UCB and MCTopM-kl-UCB. Our improved analysis also applies to RhoRand. Explicit expressions for $C_{\boldsymbol{\mu}}$, $D_{\boldsymbol{\mu}}$ can be found in the proof given in Appendix D.1.

Lemma 8. *For any $\boldsymbol{\mu} \in \mathcal{P}_M$, if player $j \in \{1, \dots, M\}$ uses the* RandTopM-, MCTopM- *or* RhoRand-kl-UCB *decentralized policy with exploration function $f(t) = \log(t) + 3\log\log(t)$, then for any sub-optimal arm $k \in M$-worst, there exists two problem depend constants $C_{\boldsymbol{\mu}}$, $D_{\boldsymbol{\mu}}$ such that*

$$\mathbb{E}_{\mu}[T_k^j(T)] \leq \frac{\log(T)}{\mathrm{kl}(\mu_k, \mu_M^*)} + C_{\boldsymbol{\mu}}\sqrt{\log(T)} + D_{\boldsymbol{\mu}}\log\log(T) + 3M + 1. \tag{18}$$

It is important to notice that the leading constant in front of $\log(T)$ is the same as in the constant featured in Equation (7) of Theorem 6. This result proves that the lower bound on sub-optimal selections is asymptotically matched for the three considered algorithms. This is a strong improvement in comparison to the previous state-of-the-art results (Liu and Zhao, 2010; Anandkumar et al., 2011).

As announced, Lemma 9 controls the number of switches of arm that are due to the current arm leaving $\widehat{M}^j(t)$, for both RandTopM and MCTopM. It essentially proves that Lines 3-5 in Algorithm 1 (when a new arm is sampled from the non-empty subset of $\widehat{M}^j(t)$) happen a logarithmic number of times. The proof of this result is given in Appendix D.2.

Lemma 9. *For any $\boldsymbol{\mu} \in \mathcal{P}_M$, any player $j \in \{1, \dots, M\}$ using* RandTopM- *or* MCTopM-kl-UCB*, and any arm k, it holds that*

$$\sum_{t=1}^{T} \mathbb{P}\left(A^j(t) = k, k \notin \widehat{M}^j(t)\right) = \left(\sum_{k', \mu_{k'} < \mu_k} \frac{1}{\mathrm{kl}(\mu_k, \mu_{k'})} + \sum_{k', \mu_{k'} > \mu_k} \frac{1}{\mathrm{kl}(\mu_{k'}, \mu_k)}\right)\log(T) + o(\log(T)).$$

6.2 Regret Analysis of MCTopM-kl-UCB

For MCTopM, we are furthermore able to obtain a logarithmic regret upper bound, by proposing an original approach to control the number of collisions under this algorithm. First, we can bound the number of collisions by the number of collisions for players not yet "fixed on their arms" ($\overline{s^j(t)}$), that we can then bound by the number of changes of arms (*cf* proof in Appendix D.3). An interesting consequence of the proof of this result is that it also bounds the number of *switches of arms*, $\sum_{t=1}^{T} \mathbb{P}(A^j(t+1) \neq A^j(t))$, and this additional guarantee was never clearly stated for previous state-of-the-art works, like RhoRand. Even though minimizing switching was not a goal[4], this guarantee is interesting for Cognitive Radio applications, where switching arms means reconfiguring a radio hardware, an operation that costs energy.

Lemma 10. *For any $\boldsymbol{\mu} \in \mathcal{P}_M$, if all players use the* MCTopM-kl-UCB *decentralized policy, and $M \leq K$, then the total average number of collisions (on all arms) is upper-bounded by (Proved in Appendix D.3)*

$$\mathbb{E}_{\mu}\left[\sum_{k=1}^{K} C_k(T)\right] \leq M^2(2M+1)\left(\sum_{a,b=1,\dots,K,\ \mu_a < \mu_b} \frac{1}{\mathrm{kl}(\mu_a, \mu_b)}\right)\log(T) + o(\log T)\,. \tag{19}$$

Note that this bound is in $\mathcal{O}(M^3)$, which significantly improves the $\mathcal{O}\left(M\binom{2M-1}{M}\right)$ proved by Anandkumar et al. (2011) for RhoRand. It is worse than the $\mathcal{O}(M^2)$ proved by Rosenski et al. (2016)

4. Introducing *switching costs*, like it was done in previous works, *e.g.*, Tomer et al. (2017), could be an interesting future work.

for MusicalChair. However, unlike MusicalChair our algorithm does not need the knowledge of $\mu^*_M - \mu^*_{M+1}$.

Now that the sub-optimal arms selections and the collisions are both proved to be at most logarithmic in Lemmas 8 and 10, it follows from our regret decomposition (Lemma 3) together with Lemma 7 that the regret of MCTopM-kl-UCB is logarithmic. More precisely, one obtains a finite-time problem-depend upper bound on the regret of this algorithm.

Theorem 11. *If all M players use* MCTopM-kl-UCB*, and $M \leq K$, then for any problem $\boldsymbol{\mu} \in \mathcal{P}_M$, there exists a problem dependent constant $G_{M,\boldsymbol{\mu}}$, such that the regret satisfies:*

$$R_T(\boldsymbol{\mu}, M, \rho) \leq G_{M,\boldsymbol{\mu}} \log(T) + o(\log T)\,. \tag{20}$$

6.3 Discussion on Selfish

The analysis of Selfish is harder, but we tried our best to obtain some understanding of the behavior of this algorithm, that seems to be doing surprisingly well in many contexts, as in our experiments with $K = 9$ arms and in extensive experiments not reported in this paper. However, a disappointing result is that we found simple problems, usually with small number of arms, for which the algorithm may fail. For example with $M = 2$ or $M = 3$ players competing for $K = 3$ arms, with means $\boldsymbol{\mu} = [0.1, 0.5, 0.9]$, the histograms in Figure 8 suggests that with a small probability, the regret R_T of Selfish-kl-UCB can be very large. We provide a discussion in Appendix E about when such situations may happen, including a conjectured (constant, but small) lower bound on the probability that Selfish experience collision almost at every round. This result would then prevent Selfish from having a logarithmic regret. However, it is to be noted that the lower bound of Theorem 6 does not apply to the censored observation model (III) under which Selfish operates, and it is not known yet whether logarithmic regret is at all possible.

7. Conclusion and future work

To summarize, we presented three variants of Multi-Player Multi-Arm Bandits, with different level of feedback being available to the decentralized players, under which we proposed efficient algorithms. For the two easiest models –with sensing–, our theoretical contribution improves both the state-of-the-art upper and lower bounds on the regret. In the absence of sensing, we also provide some motivation for the practical use of the interesting Selfish heuristic, a simple index policy based on hybrid indices that are directly taking into account the collision information.

This work suggests several interesting further research directions. First, we want to investigate the notion of *optimal algorithms* in the decentralized multi-player model with sensing information. So far we provided the first matching upper and lower bound on the expected number of sub-optimal arms selections, which suggests some form of (asymptotic) optimality. However, sub-optimal draws turn out not be the dominant terms in the regret, both in our upper bounds and in practice, thus an interesting future work is to identify some notion of *minimal number of collisions*. Second, it remains an open question to know if a simple decentralized algorithm can be as efficient as MCTopM without knowing M in advance, or in dynamic settings (when M can change in time). We shall start by proposing variants of our algorithm that are inspired by the RhoEst variant of RhoRand proposed by Anandkumar et al. (2011). Finally, we want to strengthen the guarantees obtained in the absence of sensing, that is to know whether logarithmic regret is achievable and to have a better analysis of the Selfish approach. Indeed, in most cases, it performs comparably to RandTopM even with limited feedback and without knowing the number of players M, which makes it a good candidate for applications to Internet of Things networks.

Acknowledgements: This work is supported by the French Agence Nationale de la Recherche (ANR), under grant ANR-16-CE40-0002 (project BADASS), by the CNRS under the PEPS project BIO, by the French Ministry of Higher Education and Research (MENESR) and ENS Paris-Saclay. Thanks to Christophe Moy, Rémi Bonnefoi and Vincent Gouldieff for useful discussions.

References

R. Agrawal. Sample mean based index policies by $\mathcal{O}(\log n)$ regret for the Multi-Armed Bandit problem. *Advances in Applied Probability*, 27(04):1054–1078, 1995.

S. Agrawal and N. Goyal. Analysis of Thompson sampling for the Multi-Armed Bandit problem. In *JMLR, Conference On Learning Theory*, 2012.

A. Anandkumar, N. Michael, and A. K. Tang. Opportunistic Spectrum Access with multiple users: Learning under competition. In *IEEE INFOCOM*, 2010.

A. Anandkumar, N. Michael, A. K. Tang, and S. Agrawal. Distributed algorithms for learning and cognitive medium access with logarithmic regret. *IEEE Journal on Selected Areas in Communications*, 29(4):731–745, 2011.

V. Anantharam, P. Varaiya, and J. Walrand. Asymptotically efficient allocation rules for the Multi-Armed Bandit problem with multiple plays - Part I: IID rewards. *IEEE Transactions on Automatic Control*, 32 (11):968–976, 1987.

P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time Analysis of the Multi-armed Bandit Problem. *Machine Learning*, 47(2):235–256, 2002a.

P. Auer, N. Cesa-Bianchi, Y. Freund, and R. Schapire. The Non-Stochastic Multi Armed Bandit Problem. *SIAM Journal of Computing*, 32(1):48–77, 2002b.

O. Avner and S. Mannor. Learning to Coordinate Without Communication in Multi-User Multi-Armed Bandit Problems. *arXiv preprint arXiv:1504.08167*, 2015.

O. Avner and S. Mannor. Multi-user lax communications: a Multi-Armed Bandit approach. In *IEEE INFOCOM*. IEEE, 2016.

R. Bonnefoi, L. Besson, C. Moy, E. Kaufmann, and J. Palicot. Multi-Armed Bandit Learning in IoT Networks: Learning helps even in non-stationary settings. In *12th EAI Conference on Cognitive Radio Oriented Wireless Network and Communication*, CROWNCOM Proceedings, 2017.

S. Bubeck, N. Cesa-Bianchi, et al. Regret Analysis of Stochastic and Non-Stochastic Multi-Armed Bandit Problems. *Foundations and Trends® in Machine Learning*, 5(1), 2012.

O. Cappé, A. Garivier, O-A. Maillard, R. Munos, and G. Stoltz. Kullback-Leibler upper confidence bounds for optimal sequential allocation. *Annals of Statistics*, 41(3):1516–1541, 2013.

O. Chapelle, E. Manavoglu, and R. Rosales. Simple and scalable response prediction for display advertising. *Transactions on Intelligent Systems and Technology*, 2014.

A. Garivier, P. Ménard, and G. Stoltz. Explore First, Exploit Next: The True Shape of Regret in Bandit Problems. *arXiv preprint arXiv:1602.07182*, 2016.

W. Jouini, D. Ernst, C. Moy, and J. Palicot. Multi-Armed Bandit based policies for Cognitive Radio's decision making issues. In *International Conference Signals, Circuits and Systems*. IEEE, 2009.

W. Jouini, D. Ernst, C. Moy, and J. Palicot. Upper Confidence Bound Based Decision Making Strategies and Dynamic Spectrum Access. In *IEEE International Conference on Communications*, 2010.

D. Kalathil, N. Nayyar, and R. Jain. Decentralized Learning for Multi-Player Multi-Armed Bandits. In *IEEE Conference on Decision and Control*, 2012.

E. Kaufmann, O. Cappé, and A. Garivier. On Bayesian Upper Confidence Bounds for Bandit Problems. In *AISTATS*, pages 592–600, 2012a.

E. Kaufmann, N. Korda, and R. Munos. *Thompson Sampling: an Asymptotically Optimal Finite-Time Analysis*, pages 199–213. Springer, Berlin Heidelberg, 2012b.

J. Komiyama, J. Honda, and H. Nakagawa. Optimal Regret Analysis of Thompson Sampling in Stochastic Multi-Armed Bandit Problem with Multiple Plays. In *International Conference on Machine Learning*, volume 37, pages 1152–1161, 2015.

T. L. Lai and H. Robbins. Asymptotically Efficient Adaptive Allocation Rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985.

L. Li, W. Chu, J. Langford, and R. E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World Wide Web*, pages 661–670. ACM, 2010.

K. Liu and Q. Zhao. Distributed learning in Multi-Armed Bandit with multiple players. *IEEE Transaction on Signal Processing*, 58(11):5667–5681, 2010.

J. Mitola and G. Q. Maguire. Cognitive Radio: making software radios more personal. *IEEE Personal Communications*, 6(4):13–18, 1999.

H. Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–535, 1952.

J. Rosenski, O. Shamir, and L. Szlak. Multi-Player Bandits – A Musical Chairs Approach. In *International Conference on Machine Learning*, pages 155–163, 2016.

C. Tekin and M. Liu. Online Learning in Decentralized Multi-User Spectrum Access with Synchronized Explorations. In *IEEE Military Communications Conference*, 2012.

W. R. Thompson. On the Likelihood that One Unknown Probability Exceeds Another in View of the Evidence of Two Samples. *Biometrika*, 25, 1933.

K. Tomer, L. Roi, and Y. Mansour. Bandits with Movement Costs and Adaptive Pricing. In *30th Annual Conference on Learning Theory (COLT)*, volume 65 of *JMLR Workshop and Conference Proceedings*, pages 1242–1268, 2017.

Appendix A. Regret Decompositions

A.1 Proof of Lemma 3

Using the definition of regret R_T from (3), and this collision indicator $\eta^j(t) \coloneqq \mathbb{1}(\overline{C^j(t)})$,

$$R(T) = \left(\sum_{k=1}^{M} \mu_k^*\right) T - \mathbb{E}_\mu\left[\sum_{t=1}^{T}\sum_{j=1}^{M} Y_{A^j(t),t}\eta^j(t)\right] = \left(\sum_{k=1}^{M} \mu_k^*\right) T - \mathbb{E}_\mu\left[\sum_{t=1}^{T}\sum_{j=1}^{M} \mu_{A^j(t)}\eta^j(t)\right]$$

The last equality comes from the linearity of expectations, and the fact that $\mathbb{E}_\mu[Y_{k,t}] = \mu_k$ (for all t, from the *i.i.d.* hypothesis), and the independence from $A^j(t)$, $\eta^j(t)$ and $Y_{k,t}$ (observed *after* playing $A^j(t)$). So $\mathbb{E}_\mu[Y_{A^j(t),t}\eta^j(t)] = \sum_k \mathbb{E}_\mu[\mu_k \mathbb{1}(A^j(t),t)\eta^j(t)] = \mathbb{E}_\mu[\mu_{A^j(t)}\eta^j(t)]$. And so

$$\begin{aligned} R(T) &= \mathbb{E}_\mu\left[\sum_{t=1}^{T}\sum_{j\in M\text{-best}} \mu_j - \sum_{t=1}^{T}\sum_{j=1}^{M} \mu_{A^j(t)}\eta^j(t)\right] \\ &= \left(\frac{1}{M}\sum_{j\in M\text{-best}} \mu_j\right) - \sum_{k=1}^{K}\sum_{j=1}^{M} \mu_k \mathbb{E}_\mu\left[T_k^j(T)\right] + \sum_{k=1}^{K} \mu_k \mathbb{E}_\mu\left[\mathcal{C}_k(T)\right]. \end{aligned}$$

For the first term, we have $TM = \sum_{k=1}^{K}\sum_{j=1}^{M} \mathbb{E}_\mu\left[T_k^j(T)\right]$, and if we denote $\overline{\mu}^* \coloneqq \frac{1}{M}\sum_{j\in M\text{-best}} \mu_j$ the average mean of the M-best arms, then,

$$= \sum_{k=1}^{K}\sum_{j=1}^{M} (\overline{\mu}^* - \mu_k)\mathbb{E}_\mu\left[T_k^j(T)\right] + \sum_{k=1}^{K} \mu_k \mathbb{E}_\mu\left[\mathcal{C}_k(T)\right].$$

Let $\overline{\Delta_k} \coloneqq \overline{\mu}^* - \mu_k$ be the gap between the mean of the arm k and the M-best average mean, and if M^* denotes the index of the worst of the M-best arms (*i.e.*, $M^* = \arg\min_{k\in M\text{-best}}(\mu_k)$), then by splitting $\{1,\dots,K\}$ into three disjoint sets $M\text{-best} \cup M\text{-worst} = (M\text{-best} \setminus \{M^*\}) \cup \{M^*\} \cup M\text{-worst}$, we get

$$\begin{aligned} &= \sum_{k\in M\text{-best}\setminus\{M\}} \overline{\Delta_k}\mathbb{E}_\mu\left[T_k(T)\right] + \overline{\Delta_{M^*}}\mathbb{E}_\mu\left[T_{M^*}(T)\right] \\ &\quad + \sum_{k\in M\text{-worst}} \overline{\Delta_k}\mathbb{E}_\mu\left[T_k(T)\right] + \sum_{k=1}^{K} \mu_k \mathbb{E}_\mu\left[\mathcal{C}_k(T)\right]. \end{aligned}$$

But for $k = M^*$, $T_{M^*}(T) = TM^* - \sum_{k\in M\text{-best}\setminus\{M\}} \mathbb{E}_\mu\left[T_k(T)\right] - \sum_{k\in M\text{-worst}} \mathbb{E}_\mu\left[T_k(T)\right]$, so by recombining the terms, we obtain,

$$\begin{aligned} &= \sum_{k\in M\text{-best}\setminus\{M\}} (\overline{\Delta_k} - \overline{\Delta_{M^*}})\mathbb{E}_\mu\left[T_k(T)\right] + \overline{\Delta_{M^*}} TM^* \\ &\quad + \sum_{k\in M\text{-worst}} (\overline{\Delta_k} - \overline{\Delta_{M^*}})\mathbb{E}_\mu\left[T_k(T)\right] + \sum_{k=1}^{K} \mu_k \mathbb{E}_\mu\left[\mathcal{C}_k(T)\right]. \end{aligned}$$

The term $\overline{\Delta_k} - \overline{\Delta_{M^*}}$ simplifies to $\mu_{M^*} - \mu_k$, and so $\overline{\Delta_{M^*}} = \frac{1}{M}\sum_{k=1}^{M} \mu_k - \mu_{M^*}$ by definition of $\overline{\mu}^*$. And for $k = M^*$, $\mu_{M^*} - \mu_k = 0$, so the first sum can be written for $k = 1,\dots,M$ only, so

$$\begin{aligned} R(T) = &\sum_{k\in M\text{-best}} (\mu_{M^*} - \mu_k)\mathbb{E}_\mu\left[T_k(T)\right] + \sum_{k\in M\text{-best}} (\mu_k - \mu_{M^*})T \\ &+ \sum_{k\in M\text{-worst}} (\mu_{M^*} - \mu_k)\mathbb{E}_\mu\left[T_k(T)\right] + \sum_{k=1}^{K} \mu_k \mathbb{E}_\mu\left[\mathcal{C}_k(T)\right] \end{aligned}$$

And so we obtain the decomposition with three terms (a), (b) and (c).

$$R(T) = \sum_{k \in M\text{-best}} (\mu_k - \mu_{M^*}) \left(T - \mathbb{E}_\mu \left[T_k(T)\right]\right) + \sum_{k \in M\text{-worst}} (\mu_{M^*} - \mu_k) \mathbb{E}_\mu \left[T_k(T)\right] + \sum_{k=1}^{K} \mu_k \mathbb{E}_\mu \left[C_k(T)\right].$$

A.2 Proof of Lemma 4

Note that term (c) is clearly lower bounded by 0 but it is not obvious for (b) as there is no reason for $T_k(T)$ to be upper bounded by T. Let $T_k^!(T) := \sum_{t=1}^{T} \mathbb{1}(\exists! j, A^j(t) = k)$, where the notation $\exists!$ stands for "there exists a unique". Then $T_k(T) = \sum_{t=1}^{T} \sum_{j=1}^{M} \mathbb{1}(A^j(t) = k)$ can be decomposed as

$$T_k(T) = \sum_{t=1}^{T} \mathbb{1}(\exists! j, A^j(t) = k) + \sum_{t=1}^{T} \sum_{j=1}^{M} c_{k,t} \mathbb{1}(A^j(t) = k) = T_k^!(T) + C_k(T).$$

By focusing on the two terms (b) + (c) from the decomposition of $R_T(\boldsymbol{\mu}, M, \rho)$ from Lemma 3, we have

$$\begin{aligned}
(b) + (c) &= \sum_{k \in M\text{-best}} (\mu_k - \mu_M^*)(T - \mathbb{E}_\mu[T_k^!(T)]) + \sum_{k \in M\text{-best}} \mu_M^* \mathbb{E}_\mu[C_k(T)] \\
&\quad + \sum_{k=1}^{M} \mu_k \mathbb{E}_\mu[C_k(T)] - \sum_{k \in M\text{-best}} \mu_k \mathbb{E}_\mu[C_k(T)] \\
&= \sum_{k \in M\text{-best}} (\mu_k - \mu_M^*)(T - \mathbb{E}_\mu[T_k^!(T)]) + \sum_{k \in M\text{-best}} \mu_M^* \mathbb{E}_\mu[C_k(T)] + \sum_{k \in M\text{-worst}} \mu_k \mathbb{E}_\mu[C_k(T)] \\
&= \sum_{k \in M\text{-best}} (\mu_k - \mu_M^*)(T - \mathbb{E}_\mu[T_k^!(T)]) + \sum_{k=1}^{M} \min(\mu_M^*, \mu_k) \mathbb{E}_\mu[C_k(T)].
\end{aligned}$$

And now both terms are non-negative, as $T_k^!(T) \leq T$, $\min(\mu_M^*, \mu_k) \geq 0$, and $C_k(T) \geq 0$, so (b) + (c) ≥ 0 which proves that $R_T(\boldsymbol{\mu}, M, \rho) \geq$ (a), as wanted.

A.3 Proof of Lemma 7

Recall that we want to upper bound $(b) := \sum_{k \in M\text{-best}} (\mu_k - \mu_{M*}) \left(T - \mathbb{E}_\mu[T_k(T)]\right)$. First, we observe that, for all $k \in M$-best,

$$T - \mathbb{E}_\mu[T_k(T)] \quad \leq \quad T - \mathbb{E}_\mu\left[\sum_{t=1}^{T} \mathbb{1}(\exists j : A^j(t) = k)\right] = \mathbb{E}_\mu\left[\sum_{t=1}^{T} \mathbb{1}(\forall j, A_j(t) \neq k)\right] = \mathbb{E}_\mu\left[\sum_{t=1}^{T} \mathbb{1}(k \notin \widehat{S}_t)\right],$$

where we denote by $\widehat{S}_t = \{A^j(t), j \in \{1, \dots, M\}\}$ the set of selected arms at time t (with no repetition). With this notation one can write

$$\begin{aligned}
(b) \quad &\leq \quad (\mu_1 - \mu_{M^*}) \sum_{k \in M\text{-best}} (T - \mathbb{E}_\mu[T_k(T)]) \leq (\mu_1 - \mu_{M^*}) \mathbb{E}_\mu\left[\sum_{k \in M\text{-best}} \sum_{t=1}^{T} \mathbb{1}(k \notin \widehat{S}_t)\right] \\
&= \quad (\mu_1 - \mu_{M^*}) \mathbb{E}_\mu\left[\sum_{t=1}^{T} \sum_{k \in M\text{-best}} \mathbb{1}(k \notin \widehat{S}_t)\right].
\end{aligned}$$

The quantity $\sum_{k \in M\text{-best}} \mathbb{1}(k \notin \widehat{S}_t)$ counts the number of optimal arms that have not been selected at time t. For each mis-selection of an optimal arm, there either exists a sub-optimal arm that has been selected, or an arm in M-best on which a collision occurs. Hence

$$\sum_{k \in M\text{-best}} \mathbb{1}(k \notin \widehat{S}_t) = \sum_{k \in M\text{-best}} \mathbb{1}(C_k(t)) + \sum_{k \in M\text{-worst}} \mathbb{1}(\exists j : A^j(t) = k),$$

which yields

$$\mathbb{E}_\mu\left[\sum_{t=1}^{T}\sum_{k\in M\text{-best}}\mathbb{1}(k\notin\widehat{S}_t)\right]\leq\sum_{k\in M\text{-best}}\mathbb{E}_\mu\left[\mathcal{C}_k(T)\right]+\sum_{k\in M\text{-worst}}\mathbb{E}_\mu\left[T_k(T)\right]$$

and Lemma 7 follows.

Appendix B. Lower Bound: Proof of Theorem 6

B.1 Proof of Theorem 6

The lower bound that we present relies on the following *change-of-distribution* lemma that we prove in the next section, following recent arguments from Garivier et al. (2016) that have to be adapted to incorporate the collision information.

Lemma 12. *Under observation model (I) and (II), for every event A that is $\mathcal{F}_T^j$-measurable, considering two multi-player bandit models denoted by μ and λ respectively, it holds that*

$$\sum_{k=1}^{K}\mathbb{E}_\mu\left[T_k^j(T)\right]\mathrm{kl}(\mu_k,\lambda_k)\geq\mathrm{kl}\left(\mathbb{P}_\mu(A),\mathbb{P}_\lambda(A)\right). \tag{21}$$

Let k be a sub-optimal arm under μ, fix $\varepsilon\in(0,\mu^*_{M-1}-\mu^*_M)$, and let λ be the bandit instance such that

$$\left\{\begin{array}{rcl}\lambda_\ell&=&\mu_\ell\quad\text{for all }\ell\neq k,\\ \lambda_k&=&\mu_{M^*}+\varepsilon.\end{array}\right.$$

Clearly, $\lambda\in\mathcal{P}_M$ also, and the set of M best arms under μ and λ differ by one arm: if $M\text{-best}_\mu=\{1^*,\dots,M^*\}$ then $M\text{-best}_\lambda=\{1^*,\dots,(M-1)^*,k\}$. Thus, one expects the ($\mathcal{F}_T^j$-mesurable) event

$$A_T=\left(T_k^j(T)>\frac{T}{2M}\right)$$

to have a small probability under μ (under which k is sub-optimal) and a large probability under λ (under which k is one of the optimal arms, and is likely to be drawn a lot).

Applying the inequality in Lemma 12, and noting that the sum in the left-hand side reduces to on term as there is a single arm whose distribution is changed, one obtains

$$\begin{aligned}\mathbb{E}_\mu\left[T_k^j(T)\right]\mathrm{kl}(\mu_k,\mu_{M^*}+\varepsilon)&\geq\mathrm{kl}\left(\mathbb{P}_\mu(A_T),\mathbb{P}_\lambda(A_T)\right),\\ \mathbb{E}_\mu\left[T_k^j(T)\right]\mathrm{kl}(\mu_k,\mu_{M^*}+\varepsilon)&\geq(1-\mathbb{P}_\mu(A_T))\log\left(\frac{1}{\mathbb{P}_\lambda(\overline{A_T})}\right)-\log(2),\end{aligned}$$

using the fact that the binary KL-divergence satisfies $\mathrm{kl}(x,y)=\mathrm{kl}(1-x,1-y)$ as well as the inequality $\mathrm{kl}(x,y)\geq x\log(1/y)-\log(2)$, proved by Garivier et al. (2016). Now, using Markov inequality yields

$$\begin{aligned}\mathbb{P}_\mu\left(A_T\right)&\leq 2M\frac{\mathbb{E}_\mu\left[T_k^j(T)\right]}{T}=:x_T,\\ \mathbb{P}_\lambda\left(\overline{A_T}\right)&=\mathbb{P}_\lambda\left(\frac{T}{M}-T_k^j(T)>\frac{T}{2M}\right)\leq 2M\frac{\mathbb{E}_\lambda\left[\frac{T}{M}-T_k^j(T)\right]}{T}=:\frac{y_T}{T},\end{aligned}$$

which defines two sequences x_T and y_T, such that

$$\mathbb{E}_\mu\left[T_k^j(T)\right]\mathrm{kl}(\mu_k, \mu_{M^*}+\varepsilon) \quad \geq \quad (1-x_T)\log\left(\frac{T}{y_T}\right) - \log(2). \tag{22}$$

The strong uniform efficiency assumption (see Definition 5) further tells us that $x_T \to 0$ (as $\mathbb{E}_\mu[T_k^j(T)] = o(T^\alpha)$ for all α) and $y_T = o(T^\alpha)$ when $T \to \infty$, for all $\alpha \in (0,1)$. As a consequence, observe that $\log(y_T)/\log(T) \to 0$ when T tends to infinity and

$$\frac{(1-x_T)\log(T/y_T) - \log(2)}{\log(T)} = 1 - x_T - (1-x_T)\frac{\log(y_T)}{\log(T)} - \frac{\log(2)}{\log(T)}$$

tends to one when T tends to infinity. From Equation (22), this yields

$$\liminf_{T\to\infty} \frac{\mathbb{E}_\mu[T_k^j(T)]}{\log(T)} \geq \frac{1}{\mathrm{kl}(\mu_k, \mu_{M^*}+\varepsilon)}, \tag{23}$$

for all $\varepsilon \in (0, \mu^*_{M-1} - \mu^*_M)$. Letting ε go to zero gives the conclusion (as kl is continuous).

B.2 Proof of Lemma 12

Under observation model (I) and (II), the strategy $A^j(t)$ decides which arm to play based on the information contained in I_{t-1}, where $I_0 = U_0$ and

$$\forall t > 0, \quad I_t = (U_0, Y_1, C_1, U_1, \ldots, Y_t, C_t, U_t)$$

where $Y_t := Y_{A^j(t-1),t}$ denotes the sensing information, $C_t := C^j(t)$ denotes the collision information (not always completely exploited under observation model (II)) and U_t denotes some external source of randomness[5] useful to select $A^j(t)$. Formally, one can say that $A^j(t)$ is $\sigma(I_{t-1})$ measurable[6] (as $\mathcal{F}^j(t) \subseteq \sigma(I_t)$, with an equality under observation model (I)).

Under two bandit models μ and λ, we let $\mathbb{P}_\mu^{I_t}$ (resp. $\mathbb{P}_\lambda^{I_t}$) be the distribution of the observations under model μ (resp. λ), given a fixed algorithm. Using the exact same technique as Garivier et al. (2016) (the contraction of entropy principle), one can establish that for any event A that is $\sigma(I_t)$-measurable[7],

$$\mathrm{KL}\left(\mathbb{P}_\mu^{I_t}, \mathbb{P}_\lambda^{I_t}\right) \geq \mathrm{kl}\left(\mathbb{P}_\mu(A), \mathbb{P}_\lambda(A)\right).$$

The next step is to relate the complicated KL-divergence $\mathrm{KL}\left(\mathbb{P}_\mu^{I_t}, \mathbb{P}_\lambda^{I_t}\right)$ to the number of arm selections. Proceeding similarly as Garivier et al. (2016), one can write, using the chain rule for KL-divergence, that

$$\mathrm{KL}\left(\mathbb{P}_\mu^{I_t}, \mathbb{P}_\lambda^{I_t}\right) = \mathrm{KL}\left(\mathbb{P}_\mu^{I_{t-1}}, \mathbb{P}_\lambda^{I_{t-1}}\right) + \mathrm{KL}\left(\mathbb{P}_\mu^{Y_t,C_t,U_t|I_{t-1}}, \mathbb{P}_\lambda^{Y_t,C_t,U_t|I_{t-1}}\right). \tag{24}$$

Now observe that conditionally to I_{t-1}, U_t, Y_t and C_t are independent, as once the selected arm is known, the value of the sensing Y_t does not influence the other players selecting that arm, and U_t is some exogenous randomness. Using further that the distribution of U_t is the same under μ and λ, one obtains

$$\mathrm{KL}\left(\mathbb{P}_\mu^{Y_t,C_t|I_{t-1}}, \mathbb{P}_\lambda^{Y_t,C_t|I_{t-1}}\right) = \mathrm{KL}\left(\mathbb{P}_\mu^{Y_t|I_{t-1}}, \mathbb{P}_\lambda^{Y_t|I_{t-1}}\right) + \mathrm{KL}\left(\mathbb{P}_\mu^{C_t|I_{t-1}}, \mathbb{P}_\lambda^{C_t|I_{t-1}}\right). \tag{25}$$

5. For instance, MCTopM, RandTopM and RhoRand draws from a uniform variable in $\{1, \ldots, M\}$ for new ranks or arms.
6. $\sigma(I_t)$ denotes the sigma-Algebra generated by the observations I_t.
7. In the work of Garivier et al. (2016), the statement is more general and the probability of an event A is replaced by the expectation of any $\mathcal{F}_T$-measurable random variable Z bounded in $[0,1]$.

The first term in (25) can be rewritten using the same argument as Garivier et al. (2016), that relies on the fact that conditionally to I_{t-1}, Y_t is a Bernoulli distribution with mean $\mu_{A^j(t)}$ under the instance μ and $\lambda_{A^j(t)}$ under the instance λ:

$$\begin{aligned}\mathrm{KL}\left(\mathbb{P}_\mu^{Y_t|I_{t-1}}, \mathbb{P}_\lambda^{Y_t|I_{t-1}}\right) &= \mathbb{E}_\mu\left[\mathbb{E}_\mu\left[\mathrm{KL}\left(\mathbb{P}_\mu^{Y_t|I_{t-1}}, \mathbb{P}_\lambda^{Y_t|I_{t-1}}\right)|I_{t-1}\right]\right] \\ &= \mathbb{E}_\mu\left[\mathrm{kl}\left(\mu_{A^j(t)}, \lambda_{A^j(t)}\right)\right] \\ &= \mathbb{E}_\mu\left[\sum_{k=1}^K \mathbb{1}(A^j(t)=k)\mathrm{kl}\left(\mu_k, \lambda_k\right)\right],\end{aligned}$$

We now show that second term in (25) is zero:

$$\begin{aligned}\mathrm{KL}\left(\mathbb{P}_\mu^{C_t|I_{t-1}}, \mathbb{P}_\lambda^{C_t|I_{t-1}}\right) &= \mathbb{E}_\mu\left[\mathbb{E}_\mu\left[\mathrm{KL}\left(\mathbb{P}_\mu^{C_t|I_{t-1}}, \mathbb{P}_\lambda^{C_t|I_{t-1}}\right)|I_{t-1}\right]\right] \\ &= \mathbb{E}_\mu\left[\mathbb{E}_\mu\left[\mathbb{E}_\mu\left[\mathrm{KL}\left(\mathbb{P}_\mu^{C_t|I_{t-1}}, \mathbb{P}_\lambda^{C_t|I_{t-1}}\right)|\bigcup_{j'\neq j} I_{t-1}^{j'}\right]|I_{t-1}\right]\right],\end{aligned}$$

where $I_{t-1}^{j'}$ denote the information available to player $j' \neq j$. Knowing the information available to all other players player C_t is an almost surely constant random variable, whose distribution is the same under μ and λ. Hence the inner expectation is zero and so does $\mathrm{KL}\left(\mathbb{P}_\mu^{C_t|I_{t-1}}, \mathbb{P}_\lambda^{C_t|I_{t-1}}\right)$.

Putting things together, we showed that

$$\mathrm{KL}\left(\mathbb{P}_\mu^{I_t}, \mathbb{P}_\lambda^{I_t}\right) = \mathrm{KL}\left(\mathbb{P}_\mu^{I_{t-1}}, \mathbb{P}_\lambda^{I_{t-1}}\right) + \mathbb{E}_\mu\left[\sum_{k=1}^K \mathbb{1}(A^j(t)=k)\mathrm{kl}\left(\mu_k, \lambda_k\right)\right].$$

Iterating this equality and using that $\mathrm{KL}\left(\mathbb{P}_\mu^{I_0}, \mathbb{P}_\lambda^{I_0}\right) = 0$ yields that

$$\sum_{k=1}^K \mathbb{E}_\mu\left[T_k^j(T)\right]\mathrm{kl}(\mu_k, \lambda_k) \geq \mathrm{kl}\left(\mathbb{P}_\mu(A), \mathbb{P}_\lambda(A)\right),$$

for all $A \in \sigma(I_T)$, in particular for all $A \in \mathcal{F}_T^j$.

Appendix C. The RandTopM algorithm

We now state precisely the RandTopM algorithm below in Algorithm 2 (page 21). It is essentially the same algorithm as MCTopM, but in a simpler version as the "Chair" aspect is removed, that is, there is no notion of state $s^j(t)$ (*cf* Algorithm 1). Player j is always considered "not fixed", and a *collision always forces a uniform sampling of the next arm* from $\widehat{M^j}(t)$ in the case of RandTopM.

Appendix D. Proofs Elements Related to Regret Upper Bounds

This Appendix includes the main proofs, missing from the content of the article, that yield the regret upper bound. We start by controlling the sub-optimal draws when the kl-UCB indices are used (instead of UCB_1), with any of our proposed algorithms (MCTopM, RandTopM) or RhoRand. Then we focus on controlling collisions for MCTopM-kl-UCB.

Let $A^j(1) \sim \mathcal{U}(\{1,\dots,K\})$ and $C^j(1) = \text{False}$
for $t = 0,\dots,T-1$ **do**
  if $A^j(t) \notin \widehat{M^j}(t)$ **then**
    if $C^j(t)$ **then** // collision
      $A^j(t+1) \sim \mathcal{U}\left(\widehat{M^j}(t)\right)$ // randomly switch
    else // randomly switch on an arm that had smaller UCB at $t-1$
      $A^j(t+1) \sim \mathcal{U}\left(\widehat{M^j}(t) \cap \left\{k : g^j_k(t-1) \leq g^j_{A^j(t)}(t-1)\right\}\right)$
    end
  else
    $A^j(t+1) = A^j(t)$ // stays on the same arm
  end
  Play arm $A^j(t+1)$, get new observations (sensing and collision),
  Compute the indices $g^j_k(t+1)$ and set $\widehat{M^j}(t+1)$ for next step.
end

Algorithm 2: The RandTopM decentralized learning policy (for a fixed underlying index policy g^j).

D.1 Control of the Sub-optimal Draws for kl-UCB: Proof of Lemma 8

Fix $k \in M$-worst and a player $j \in \{1,\dots,M\}$. The key observation is that for MCTopM, RandTopM as well as the RhoRand algorithm, it holds that

$$\left(A^j(t) = k\right) = \left(A^j(t) = k, \exists m \in M\text{-best} : g^j_m(t) < g^j_k(t)\right). \tag{26}$$

Indeed, for the three algorithms, an arm selected at time $t+1$ belongs to the set $\widehat{M^j}(t)$ of arms with M largest indices. If the sub-optimal arm k is selected at time t, it implies that $k \in \widehat{M^j}(t)$, and, because there are M arms in both M-best and $\widehat{M^j}(t)$, one of the arms in M-best must be excluded from $\widehat{M^j}(t)$. In particular, the index of arm k must be larger than the index of this particular arm m.

Using (26), one can then upper bound the number of selections of arm k by user j up to round T as

$$\begin{aligned}
\mathbb{E}_\mu[T^j_k(T)] &= \mathbb{E}_\mu\left[\sum_{t=1}^T \mathbb{1}\left(A^j(t) = k\right)\right] = \sum_{t=1}^T \mathbb{P}\left(A^j(t) = k\right). \\
&= \sum_{t=1}^T \mathbb{P}\left(A^j(t) = k,\ \ \exists m \in \{1,\dots,M\} :\ g^j_{m^*}(t) < g^j_k(t)\right).
\end{aligned}$$

Considering the relative position of the upper-confidence bound $g^j_{m^*}(t)$ and the corresponding mean $\mu^*_m = \mu_{m^*}$, one can write the decomposition

$$\begin{aligned}
\mathbb{E}_\mu[T^j_k(T)] &\leq \sum_{t=1}^T \mathbb{P}\left(A^j(t) = k,\ \ \exists m \in \{1,\dots,M\} :\ g_{m^*}(t) \leq g_k(t), \forall m \in \{1,\dots,M\} :\ g_{m^*}(t) \geq \mu^*_m\right) \\
&\quad + \sum_{t=1}^T \mathbb{P}\left(\exists m \in \{1,\dots,M\} :\ g_{m^*}(t) < \mu^*_m\right) \\
&\leq \sum_{t=1}^T \mathbb{P}\left(A^j(t) = k,\ \ \exists m \in \{1,\dots,M\} :\ \mu^*_m \leq g_k(t)\right) + \sum_{m=1}^M \sum_{t=1}^T \mathbb{P}\left(g_{m^*}(t) < \mu^*_m\right) \\
&\leq \sum_{t=1}^T \mathbb{P}\left(A^j(t) = k,\ \mu_{M^*} \leq g_k(t)\right) + \sum_{m=1}^M \sum_{t=1}^T \mathbb{P}\left(g_{m^*}(t) < \mu^*_m\right),
\end{aligned}$$

where the last inequality (for the first term) comes from the fact that μ_{M^*} is the smallest of the μ_{m^*} for $m \in \{1, \dots, M\}$.

Now each of the two terms in the right hand side can directly be upper bounded using tools developed by Cappé et al. (2013) for the analysis of kl-UCB. The rightmost term can be controlled using Lemma 13 below that relies on a self-normalized deviation inequality, whose proof exactly follows from the proof of Fact 1 in Appendix A of Cappé et al. (2013). The leftmost term can be controlled using Lemma 14 stated below, that is a direct consequence of the proof of Fact 2 in Appendix A of Cappé et al. (2013).

Lemma 13. *For any arm k, if $g_k^j(t)$ is the* kl-UCB *index with exploration function $f(t) = \log(t) + 3\log\log(t)$,*

$$\sum_{t=1}^{T} \mathbb{P}\left(g_k^j(t) < \mu_k\right) \le 3 + 4e \log\log(T). \tag{27}$$

Denote $\mathrm{kl}'(x, y)$ the derivative of the function $x \mapsto \mathrm{kl}(x, y)$ (for any fixed $y \neq 0, 1$).

Lemma 14. *For any arms k and k' such that $\mu_{k'} > \mu_k$, if $g_k^j(t)$ is the kl-UCB index with exploration function $f(t)$,*

$$\sum_{t=1}^{T} \mathbb{P}\left(A^j(t) = k, \mu_{k'} \le g_k^j(t)\right) \le \frac{f(T)}{\mathrm{kl}(\mu_k, \mu_{k'})} + \sqrt{2\pi}\sqrt{\frac{\mathrm{kl}'(\mu_k, \mu_{k'})^2}{\mathrm{kl}(\mu_k, \mu_{k'})^3}}\sqrt{f(T)} + 2\left(\frac{\mathrm{kl}'(\mu_k, \mu_{k'})}{\mathrm{kl}(\mu_k, \mu_{k'})}\right)^2 + 1.$$

Putting things together, one obtains the non-asymptotic upper bound

$$\begin{aligned}\mathbb{E}_\mu\left[T_k^j(T)\right] \le \frac{\log(T) + 3\log\log(T)}{\mathrm{kl}(\mu_k, \mu_{M^*})} + \sqrt{2\pi}\sqrt{\frac{\mathrm{kl}'(\mu_k, \mu_{M^*})^2}{\mathrm{kl}(\mu_k, \mu_{M*})^3}}\sqrt{\log(T) + 3\log\log(T)} \\ + 2\left(\frac{\mathrm{kl}'(\mu_k, \mu_{M^*})}{\mathrm{kl}(\mu_k, \mu_{M^*})}\right)^2 + 4Me\log\log(T) + 3M + 1,\end{aligned} \tag{28}$$

which yields Lemma 8, with explicit constants $C_{\boldsymbol{\mu}}$ and $D_{\boldsymbol{\mu}}$.

D.2 Proof of Lemma 9

Using the behavior of the algorithm when the current arm leaves the set $\widehat{M^j}$ (Line 4), one has

$$\begin{aligned}&\sum_{t=1}^{T} \mathbb{P}\left(A^j(t) = k, k \notin \hat{M}^j(t)\right) \\ &\le \sum_{t=1}^{T} \mathbb{P}\left(A^j(t) = k, k \notin \hat{M}^j(t), A^j(t+1) \in \widehat{M^j}(t) \cap \{k' : g_{k'}^j(t-1) \le g_k^j(t-1)\}\right) \\ &\le \sum_{t=1}^{T} \sum_{k' \neq k} \mathbb{P}\left(A^j(t) = k, A^j(t+1) = k', g_{k'}^j(t) \ge g_k^j(t), g_{k'}^j(t-1) \le g_k^j(t-1)\right) \\ &= \sum_{k' \neq k} \underbrace{\sum_{t=1}^{T} \mathbb{P}\left(A^j(t) = k, A^j(t+1) = k', g_{k'}^j(t) \ge g_k^j(t), g_{k'}^j(t-1) \le g_k^j(t-1)\right)}_{:= T_{k'}}\end{aligned}$$

Now, to control $T_{k'}$, we distinguish two cases. If $\mu_k < \mu_{k'}$, one can write

$$T_{k'} \leq \sum_{t=1}^{T} \mathbb{P}\left(g_{k'}^j(t) \leq \mu_{k'}\right) + \sum_{t=1}^{T} \mathbb{P}\left(A^j(t) = k, g_k^j(t-1) \geq \mu_{k'}\right)$$

The first term in the right hand side is $o(\log(T))$ by Lemma 13. To control the second term, we apply the same trick that led to the proof of Lemma 14 in Cappé et al. (2013). Letting $\mathrm{kl}^+(x,y) := \mathrm{kl}(x,y)\mathbb{1}(x \geq y)$, and $\widehat{\mu}_{k,s}^j$ be the empirical mean of the s first observations from arm k by player j, one has

$$\begin{aligned}
\sum_{t=1}^{T} \mathbb{P}\left(A^j(t) = k, g_k^j(t-1) \geq \mu_{k'}\right) &= \mathbb{E}\left[\sum_{t=1}^{T}\sum_{s=1}^{t-1} \mathbb{1}\left(A^j(t) = k, N_k^j(t-1) = s\right)\mathbb{1}\left(s \times \mathrm{kl}^+\left(\widehat{\mu}_{k,s}^j, \mu_k\right) \leq f(t)\right)\right] \\
&\leq \mathbb{E}\left[\sum_{s=1}^{T} \mathbb{1}\left(s \times \mathrm{kl}^+\left(\widehat{\mu}_{k,s}^j, \mu_k\right) \leq f(T)\right) \sum_{t=s-1}^{T} \mathbb{1}\left(A^j(t) = k, N_k^j(t-1) = s\right)\right] \\
&\leq \sum_{s=1}^{T} \mathbb{P}\left(s \times \mathrm{kl}^+\left(\widehat{\mu}_{k,s}^j, \mu_k\right) \leq f(T)\right), \qquad (29)
\end{aligned}$$

where the last inequality uses that for all s,

$$\sum_{t=s-1}^{T} \mathbb{1}\left(A^j(t) = k, N_k^j(t-1) = s\right) = \sum_{t=s-1}^{T} \mathbb{1}\left(A^j(t) = k, N_k^j(t) = s+1\right) \leq 1.$$

From (29), the same upper bound as that of Lemma 14 can be obtained using the tools from Cappé et al. (2013), which proves that for $T \to \infty$,

$$T_{k'} = \frac{\log(T)}{\mathrm{kl}(\mu_k, \mu_{k'})} + o(\log(T)).$$

If $\mu_k > \mu_{k'}$, we rather use that

$$T_{k'} \leq \sum_{t=1}^{T} \mathbb{P}\left(g_k^j(t) \leq \mu_k\right) + \sum_{t=1}^{T} \mathbb{P}\left(A^j(t+1) = k', g_{k'}^j(t) \geq \mu_k\right)$$

and similarly Lemma 13 and a slight variant of Lemma 14 to deal with the modified time indices yields

$$T_{k'} = \frac{\log(T)}{\mathrm{kl}(\mu_{k'}, \mu_k)} + o(\log(T)).$$

Summing over k' yields the result.

D.3 Controlling Collisions for MCTopM: Proof of Lemma 10

A key feature of both the RandTopM and MCTopM algorithms is Lemma 9, that states that the probability of switching from some arm because this arm leaves $\widehat{M}^j(t)$ is small. Its proof is postponed to the end of this section.

Figure 3 below provides a schematic representation of the execution of the MCTopM algorithm, that has to be exploited in order to properly control the number of collisions. The sketch of the proof is the following: by focusing only on collisions in the "not fixed" state, bounding the number of transitions (2) and (3) is enough. Then, we show that both the number of transitions (3) and (5) are small: as

a consequence of Lemma 9, the average number of these transitions is $\mathcal{O}(\log T)$. Finally, we use that the length of a sequence of consecutive transitions (2) is also small (on average smaller than M), and except for possibly the first one, starting a new sequence implies a previous transition (3) or (5) to arrive in the state "not fixed". This gives a logarithmic number of transitions (2) and (3), and so gives $\mathbb{E}_\mu[\sum_k C_k(T)] = \mathcal{O}(\log T)$, with explicit constants depending on $\boldsymbol{\mu}$ and M.

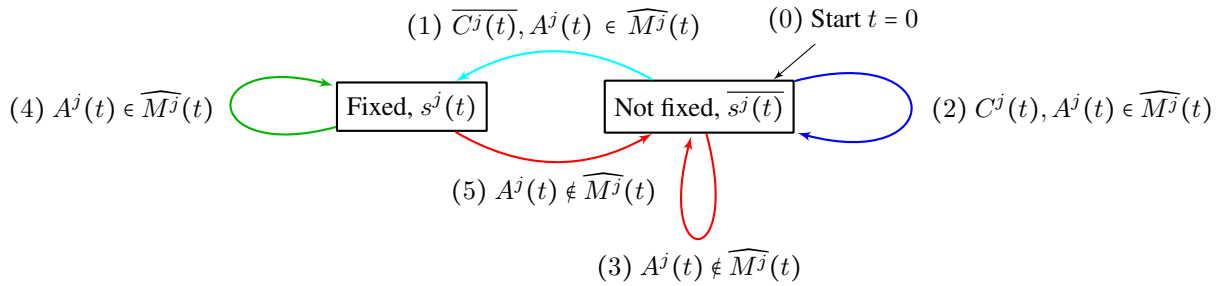


Figure 3: Player j using MCTopM, represented as "state machine" with 5 transitions. Taking one of the five transitions means playing one round of the Algorithm 1, to decide $A^j(t+1)$ using information of previous steps.

As in Algorithm 1, $s^j(t)$ is the event that player j decided to fix herself on an arm at the end of round $t-1$. Formally, $s^j(0)$ is false, and $s^j(t+1)$ is defined inductively from $s^j(t)$ as

$$s^j(t+1) = \left(s^j(t) \cup \left(\overline{s^j(t)} \cap \overline{C^j(t)}\right)\right) \cap \left(A^j(t) \in \widehat{M^j}(t)\right). \tag{30}$$

For the sake of clarity, we now explain Figure 3 in words. At step t, if player j is not fixed ($\overline{s^j(t)}$), she can have three behaviors when executing MCTopM. She keeps the same arm and goes to the other state $s^j(t)$ with transition (1), or she stays in state $\overline{s^j(t)}$, with two cases. Either she sampled $A^j(t+1)$ uniformly from $\widehat{M^j}(t) \cap \{m : g^j_m(t) \leq g^j_k(t)\}$ with transition (3), in case of collision and if $A^j(t+1) \in \widehat{M^j}(t)$, or she sampled $A^j(t+1)$ uniformly from $\widehat{M^j}(t)$ with transition (2), if $A^j(t+1) \notin \widehat{M^j}(t)$. In particular, note that if $\overline{C^j(t)}$, transition (3) is executed and not (2). Transition (3) is a uniform sampling from $\widehat{M^j}(t)$ (the "Musical Chair" step).

For player j and round t, we now introduce a few events that are useful in the proof. First, for every $x = 1, 2, 3, 4, 5$, we denote $I^j_x(t)$ the event that a transition of type (x) occurs for player j after the first t observations (*i.e.*, between round t and round $t+1$, to decide $A^j(t+1)$). Formally they are defined by

$$\begin{aligned}
I^j_1(t) &:= \left(\overline{s^j(t)}, \overline{C_j(t)}, A^j(t) \in \widehat{M^j}(t)\right), \\
I^j_2(t) &:= \left(\overline{s^j(t)}, C_j(t), A^j(t) \in \widehat{M^j}(t)\right), && \text{and} \quad I_3(t) := \left(\overline{s^j(t)}, A^j(t) \notin \widehat{M^j}(t)\right), \\
I_3(t) &:= \left(s^j(t), A^j(t) \in \widehat{M^j}(t)\right), && \text{and} \quad I_5(t) := \left(s^j(t), A^j(t) \notin \widehat{M^j}(t)\right).
\end{aligned}$$

Then, we introduce $\widetilde{C^j}(t)$ as the event that a collision occurs for player j at round t if she is not yet fixed on her arm, that is

$$\widetilde{C^j}(t) := \left(C^j(t), \overline{s^j(t)}\right). \tag{31}$$

A key observation is that $C^j(t)$ implies $\bigcup_{j'=1}^M \widetilde{C^{j'}}(t)$, as a collision necessarily involves at least one player not yet fixed on her arm ($\overline{s^{j'}(t)}$). Otherwise, if they are all fixed, *i.e.*, for all j, $s^j(t)$, then by

definition of $s^j(t)$, none of the player changed their arm from $t-1$ to t, and none experienced any collision at time $t-1$ so by induction there is no collision at time t. Thus, $\sum_{j=1}^{M} \mathbb{P}(C^j(t))$ can be upper bounded by $M \sum_{j=1}^{M} \mathbb{P}(\widetilde{C^j}(t))$ (union bound), and it follows that if $\mathcal{C}(T) := \sum_{k=1}^{K} \mathcal{C}_k(T)$ then

$$\mathbb{E}_\mu[\mathcal{C}(T)] \leq M \sum_{j=1}^{M} \sum_{t=1}^{T} \mathbb{P}(\widetilde{C^j}(t)).$$

We can further observe that $\widetilde{C^j}(t)$ implies a transition (2) or (3), as a transition (1) cannot happen in case of collision. Thus another union bound gives

$$\sum_{t=1}^{T} \mathbb{P}(\widetilde{C^j}(t)) \leq \sum_{t=1}^{T} \mathbb{P}(I_2^j(t)) + \sum_{t=1}^{T} \mathbb{P}(I_3^j(t)). \tag{32}$$

In the rest of the proof we focus on bounding the number of transitions (2) and (3).

Let $N_x^j(T)$ be the random variable denoting the number of transitions of type (x). Neglecting the event $\overline{s^j(t)}$ for $x = 3$ and $s^j(t)$ for $x = 5$, one has

$$\mathbb{E}_\mu[N_x^j(t)] = \sum_{t=1}^{T} \mathbb{P}(I_x^j(t)) \leq \sum_{t=1}^{T} \mathbb{P}\left(A^j(t) \notin \widehat{M^j}(t)\right) \leq \sum_{t=1}^{T} \sum_{k=1}^{K} \mathbb{P}\left(A^j(t) = k, k \notin \widehat{M^j}(t)\right), \tag{33}$$

which is $\mathcal{O}(\log T)$ (with known constants) by Lemma 9. In particular, this controls the second term in the right hand side of (32).

To control the first term $\sum_{t=1}^{T} \mathbb{P}(I_2^j(t))$ we introduce three sequences of random variables, the starting times $(\theta_i)_{i\geq 1}$ and the ending times $(\tau_i)_{i\geq 1}$ (possibly larger than T), of sequences during which $I_2(s)$ is true for all $s = \theta_i, \dots, \tau_i - 1$ but not before and after, that is $\forall i \in \{1, \dots, n(T)\}, \overline{I_2^j(\theta_i - 1)} \cap \bigcap_{t=\theta_i}^{\tau_i - 1} I_2^j(t) \cap \overline{I_2^j(\tau_i)}$ with $n(T)$ the number of such sequences, *i.e.*, $n(T) := \overline{\inf\{i \geq 1 : \min(\theta_i, \tau_i) \geq T\}}$ (or 0 if θ_1 does not exist). If $\theta_i = 1$, the first sequence does not have term $\overline{I_2^j(\theta_i - 1)}$.

Now we can decompose the sum on $t = 1, \dots, T$ with the use of consecutive sequences,

$$\mathbb{E}_\mu[N_2^j(t)] = \mathbb{E}_\mu\left[\sum_{t=1}^{T} \mathbb{1}\left(I_2^j(t)\right)\right] = \mathbb{E}_\mu\left[\sum_{i=1}^{n(T)} \left(\sum_{t=\theta_i}^{\tau_i - 1} 1 + \sum_{t=\tau_i}^{\theta_{i+1}-1} 0\right)\right] = \mathbb{E}_\mu\left[\sum_{i=1}^{n(T)} (\tau_i - \theta_i)\right]. \tag{34}$$

Both $n(T)$ and $\tau_i - \theta_i \geq 0$ have finite averages for any i (as $\tau_i - \theta_i \leq T$), and $n(T)$ is a *stopping time* with respect to the past events (that is, $\mathcal{F}_T^j$), and so we can obtain

$$\leq \mathbb{E}_\mu\left[n(T)\right] \times \max_{i \in \mathbb{N}} \mathbb{E}_\mu\left[\tau_i - \theta_i\right] = (\alpha) \times (\beta). \tag{35}$$

(α) To control $\mathbb{E}_\mu[n(T)]$, we can observe that the number of sequences $n(T)$ is smaller than 1 plus the number of times when *a sequence begins* (1 plus because maybe the game starts in a sequence). And beginning a sequence at time θ_i implies $\overline{I_2^j(\theta_i - 1)} \cap I_2^j(\theta_i)$, which implies a transition of type (3) or (5) at time $\theta_i - 1$, as player j is in state "not fixed" at time θ_i (transitions (1) and (4) are impossible). As stated above, $\mathbb{E}_\mu[N_x^j(T)] = \mathcal{O}(\log T)$ for both $x = 3$ and $x = 5$, and so $\mathbb{E}_\mu[n(T)] = \mathcal{O}(\log T)$ also.

(β) To control $\mathbb{E}_\mu\left[\tau_i - \theta_i\right]$, a simple argument can be used. $\bigcup_{t=\theta_i}^{\tau_i - 1} I_2^j(t)$ implies $C^j(t)$ for $\tau_i - \theta_i$ consecutive times. The very structure of RandTopM gives that in this sequence of transitions (2), the successive collisions (*i.e.*, $C^j(t-1) \cap C^j(t)$) implies that each new arm $A^j(t+1)$ for $t \in \{\theta_i, \tau_i - 1\}$ is

selected uniformly from $\widehat{M^j}(t+1)$, a set of size M with at least one available arm. Indeed, as there is $M-1$ other players, at time $t+1$ *at least* one arm in $\widehat{M^j}(t+1)$ is not selected by any player $k' \neq k$, and so player j has *at least* a probability $1/M$ to select a free arm, which implies $\overline{C^j(t+1)}$, and so implies the end of the sequence. In other words, the average length of sequences of transitions (2), $\mathbb{E}_\mu[\tau_i - \theta_i]$, is bounded by the expected number of failed trial of a repeated Bernoulli experiment, with probability of success larger than $1/M$ (by the uniform choice of $A^j(t+1)$ in a set of size M with at least one available arm). We recognize the mean of a geometric random variable, of parameter $\lambda \geq 1/M$, and so $\mathbb{E}_\mu[\tau_i - \theta_i] = \frac{1}{\lambda} \leq \frac{1}{1/M} = M$.

This finishes the proof as $\mathbb{E}_\mu[N_2^j(T)] = \sum_{t=1}^T \mathbb{P}(I_2^j(t)) = \mathcal{O}(\log T)$ and so $\sum_{t=1}^T \mathbb{P}(\widetilde{C^j}(t) \cap (A^j(t) = k)) = \mathcal{O}(\log T)$ and finally $\mathbb{E}_\mu[\mathcal{C}(T)] = \sum_{k=1}^K \mathbb{E}_\mu[\mathcal{C}^k(T)] = \mathcal{O}(\log T)$ also.

We can be more precise about the constants, all the previous arguments can be used successively:

$$\begin{aligned}
\mathbb{E}_\mu[\mathcal{C}(T)] &\leq M \sum_{j=1}^M \left(\sum_{t=1}^T \mathbb{P}(I_2^j(t)) + \sum_{t=1}^T \mathbb{P}(I_3^j(t)) \right) = M \left(\sum_{j=1}^M \mathbb{E}_\mu[N_2^j(T)] + \mathbb{E}_\mu[N_3^j(T)] \right) \qquad (36) \\
&\leq M^2 \left(\mathbb{E}_\mu[n(T)] \mathbb{E}_\mu[\theta_i - \tau_i] \right) + M^2 \mathbb{E}_\mu[N_3^1(T)] \\
&\leq M^2 (1 + \mathbb{E}_\mu[N_3^1(T)] + \mathbb{E}_\mu[N_5^1(T)]) M + M^2 \mathbb{E}_\mu[N_3^1(T)] \\
&\leq 2M^3 \mathbb{E}_\mu[N_3^1(T)] + o(\log T) + M^2 \left(\sum_{a,b=1,\dots,K,\ \mu_a < \mu_b} \frac{1}{\mathrm{kl}(\mu_a, \mu_b)} \right) \log(T) + o(\log T) \\
&\leq \left(2M^3 + M^2\right) \left(\sum_{a,b=1,\dots,K,\ \mu_a < \mu_b} \frac{1}{\mathrm{kl}(\mu_a, \mu_b)} \right) \log(T) + o(\log T) \,. \qquad (37)
\end{aligned}$$

And so we obtain the desired inequality, with explicit constants, that depend only on $\boldsymbol{\mu}$ and M.

$$\sum_{k=1}^K \mathbb{E}_\mu[\mathcal{C}^k(T)] = \mathbb{E}_\mu[\mathcal{C}(T)] \leq M^2 (2M+1) \left(\sum_{a,b=1,\dots,K,\ \mu_a < \mu_b} \frac{1}{\mathrm{kl}(\mu_a, \mu_b)} \right) \log(T) + o(\log T) \,. \qquad (38)$$

Number of switches Note that we controlled the total number of transitions (2), (3) and (5), which are the only transitions when a player can switch from arm k to arm $k' \neq k$. Thus, the total number of arm switches is also proved to be logarithmic, if all players uses the MCTopM-kl-UCB algorithm.

Strong uniform efficiency As soon as $R_T = \mathcal{O}(\log T)$ for all problem, MCTopM is clearly proved to be uniformly efficient, as $\log T$ is $o(T^\alpha)$ for any $\alpha \in (0,1)$. And as justified after Definition 5 (page 5), uniform efficiency and invariance under permutations of the users implies strong uniform efficiency, and so MCTopM satisfies Definition 5. This is a sanity check: the lower-bound of Theorem 6 indeed applies to our algorithm MCTopM, and finally this highlights that it is order-optimal for the regret, in the sense that it matches the lower-bound up-to a multiplicative constant, and optimal for the term (a).

Appendix E. Additional Discussions on Selfish

As said before, analyzing Selfish is harder, but for instance one can prove that it yields constant collisions and regret for the trivial case of $\mu_1 = \mu_2 = 1$ and $M = 2$. Empirically, when Selfish is compared to the other algorithms, it is hard to find a case when Selfish performs badly, as its (empirical average) regret always appeared logarithmic. But an issue of only visualizing the empirical average regret for a certain number of repetitions is that if a certain "bad" run happens only with small probability, it is possible that it

never happened in a simulation, or that it happened a few times but not enough to make the average regret look non-logarithmic[8]. This is why the distribution of regret, at the end of the simulations, R_T, is also displayed (in Appendix F). In a simple problem, with $M = 2$ or $M = 3$ players competing for $K = 3$ arms, for instance with means $\boldsymbol{\mu} = [0.1, 0.5, 0.9]$, the histogram in Figure 8 shows that with a small probability, the regret R_T of Selfish-kl-UCB is not small (and appears linear). Additionally, Figure 9 shows that Selfish-kl-UCB also has bad performance against random uniform problems $\boldsymbol{\mu} \in [0, 1]^K$, for $M = 2$ or 3 and $K = 3$, in a lot of cases. In comparison, the others algorithms seem to have a logarithmic regret (for $M = 2$ and all algorithms) or even a constant regret (for $M = 3$ and RandTopM and MCTopM).

The intuition behind these configurations when Selfish performs poorly is the following, if all players use the same algorithm and the same indices. If two players i and j have exactly the same vectors $[\widetilde{S_k}^i(t)] = [\widetilde{S_k}^j(t)]$ and $[T_k^i(t)] = [T_k^j(t)]$ at some time step $t \geq 1$, with different values for each k and so that the index vectors $\mathbf{g}^j(t) = \mathbf{g}^i(t)$ have different values for each arm k, then both players will take the same decisions at time t, and collide. Colliding does not change $[\widetilde{S_k}^j(t+1)]$ but increase one value in $[T_k^j(t+1)]$ by 1. Then at the next step the same conditions on $\widetilde{S}^j$ and N^j are preserved, and if the same condition on $\mathbf{g}^j(t+1)$ are also preserved, the two players will continue to collide. We did not succeed in proving mathematically that the preservation of the first hypothesis on $\widetilde{S}^j$ and N^j implies the preservation of the hypothesis on index $\mathbf{g}^j$, but numerically it turns out to be always the case: and so two players colliding in such a setting will continue to do so infinitely: we denote such configurations as *absorbing*.

We wrote a script [9] that explores formally all the possible runs, up-to a certain small time horizon, by exploring the complete *game tree*, of possible (random) rewards from the K arms and (random) actions from the M players, up-to a small depth of let say $T = 8$. Such game tree becomes quickly very large, but this was enough to confirm that with a certain small probability, function of $\mu_1, \dots, \mu_K$, two players can arrive in just a few steps in a "bad" absorbing configuration. For instance, for only $K = 2$ arms, the following game tree in Figure 4 illustrates the first 3 steps that can lead to 2 absorbing configurations. Using symbolic computations, the probability of reaching any of them was found to be $\geq \mu_1^2(1-\mu_2)^2/2 + \mu_2^2(1-\mu_1)^2/2$ for Selfish-UCB$_1$. That is far from being negligible, as it evaluates to 0.328 for $\boldsymbol{\mu} = [0.1, 0.5, 0.9]$, and a numerical simulation on 1000 runs found 325 cases of bad performance. The same game tree exploration can be made for Selfish-kl-UCB, but so far we were not able to justify why it experiences fewer cases of bad performances even though our software found the same (lower bound on) failure probability.

From the structure of such game tree, we conjecture that the probability of reaching absorbing configurations (before a certain time t) is always lower-bounded by a polynomial function of $\mu_1, \dots, \mu_K$ and $1-\mu_1, \dots, 1-\mu_K$, of degree at most t in each variable. As such, the lower bound on probability of failures should decrease when K and M increase, and this is coherent with the experiments for $K = 9$ or $K = 17$ (see Figures 13 and 14), where Selfish is shown to be uniformly more efficient than RhoRand. Of course, one cannot run an infinite number of simulations, and the smaller the probability of failure, the less likely it is to observe a failure in a finite number of runs.

Ideas to fix Selfish **?** It could be possible to change the Selfish algorithm to add a way to escape such absorbing trajectories. For instance one could imagine that after seen seeing, *e.g.*, $x = 10$ collisions in a row, a certain random action could be taken by the players. These tricks can work empirically in some

8. For instance, $0.999\log(T) + 0.001T$ looks more like $\log(T)$ than T so an event yielding linear regret with "small" probability 10^{-3} cannot be observed from a plot showing the average regret.
9. banditslilian.gforge.inria.fr/docs/complete_tree_exploration_for_MP_bandits.html

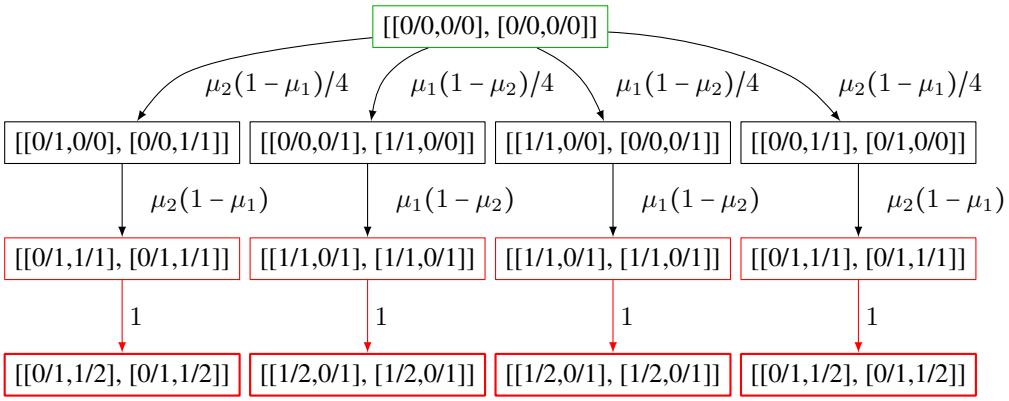


Figure 4: For $K = 2$ arms and $M = 2$ players using Selfish-UCB$_1$, for depth= 3: 2 **absorbing configurations**. Each rectangle represents a configuration, as the matrix $[[\widehat{S_k}^j(t)/T_k^j(t)]_j]_k$. Absorbing configurations from depth 2 are case of equality of the two vectors and the Selfish indices $\widetilde{g_k}^j(t)$. Transitions are labeled with their probabilities.

cases, but they are harder to analyze formally, and it is hard to tune the parameters (here x, but possibly more), and we do not find such tricks to be promising from a theoretical point-of-view.

Appendix F. Additional Figures

The plots missing from Section 5 are included here, as well as some additional numerical results.

F.1 Illustration of the lower bound

We proved in Theorem 6 that the normalized regret, *i.e.*, R_T divided by $\log T$, is asymptotically lower bounded by a constant $\mathrm{LB}(\boldsymbol{\mu}, M)$ depending on the problem $\boldsymbol{\mu}$ and the number of players M, for any ρ.

$$\liminf_{T\to+\infty} \frac{R_T(\boldsymbol{\mu}, M, \rho)}{\log T} \geq \mathrm{LB}(\boldsymbol{\mu}, M). \tag{39}$$

For an example problem with $K = 9$ arms, we display below on the x axis is the number of player, from 1 player to 9 players, and on the y axis is the value of this constant $\mathrm{LB}(\boldsymbol{\mu}, M)$, from the initial theorem and from our theorem. We chose a simple problem, with Bernoulli distributed arms, with $\boldsymbol{\mu} = [0.1, 0.2, \dots, 0.9]$. Figure 5 clearly shows that our improved lower bound is indeed larger than the initial one by Liu and Zhao (2010), and both become uninformative when $M = K$ (*i.e.*, null).

Figures 6 show the regret $R_T(\boldsymbol{\mu}, M, \rho)$ on the same example problem $\boldsymbol{\mu}$, with $K = 9$ arms and respectively $M = 6$, or 9 players, for Selfish-kl-UCB. It is just a simple way to check that the two lower bounds on the regret indeed appear as valid lower bounds empirically, and are moreover lower bounds on the count of selections ((a), displayed in cyan). The lower bounds (in black) are $C(\boldsymbol{\mu}, M) \log t$, the dashed line for Liu and Zhao's lower bound, and the continuous line is our lower bound. These plot show the regret (in red), and the three terms (a), (b), (c) in the decomposition of the regret. As explained in Lemma 3, term (b) is not always non-negative. For $M = 9$ and Selfish, (c) is actually larger than the regret, and term (a) is zero, as well as the lower bounds.

F.2 Figures from Section 5

This last Appendix includes the figures used in Section 5, with additional comments.

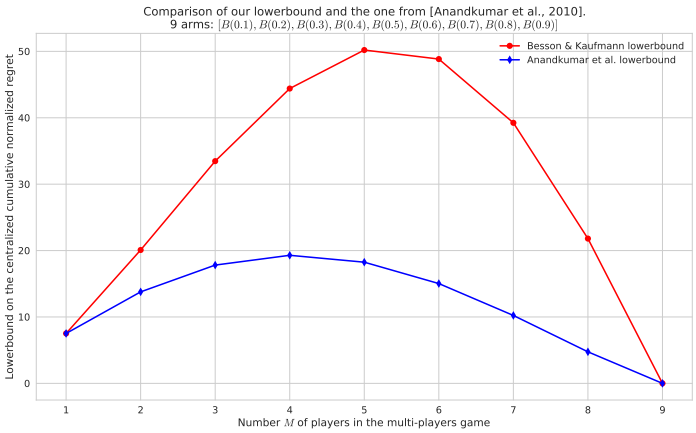
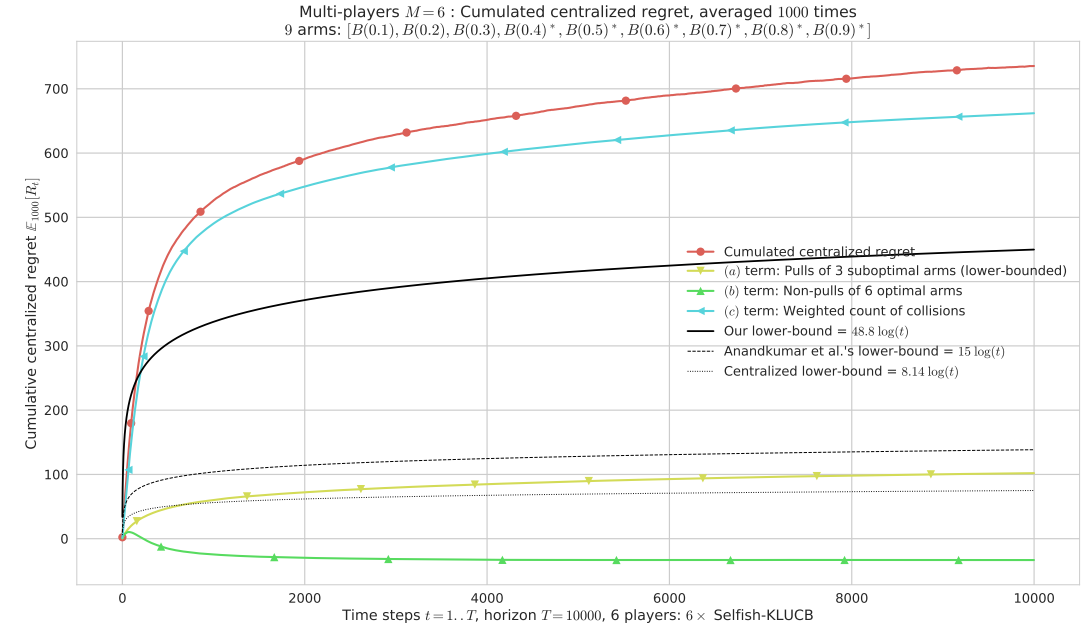


Figure 5: Comparison of our lower bound against the one from Liu and Zhao (2010), on a simple problem with 9 Bernoulli arms, of means $\boldsymbol{\mu} = [0.1, 0.2, \ldots, 0.9]$, as a function of the number of players M.

Note: the simulation code used for the experiments is using Python 3. It is open-sourced at `https://GitHub.com/SMPyBandits/SMPyBandits` and fully documented at `https://SMPyBandits.GitHub.io`.



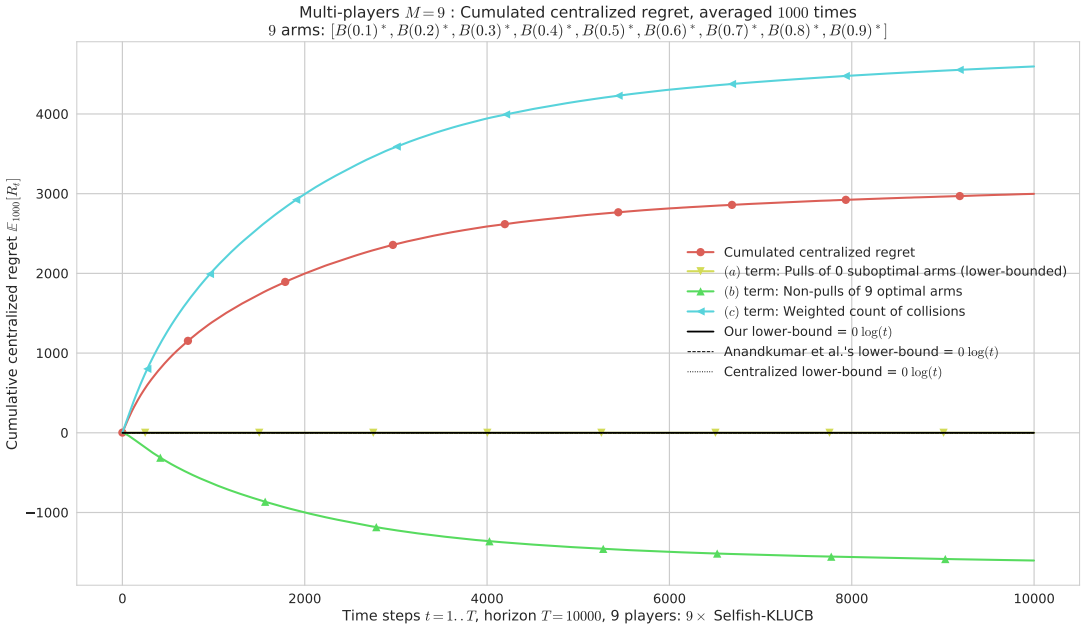


Figure 6: Regret with its three terms (a), (b), (c), and lower bounds (8) and (9) in **black**, for Selfish-kl-UCB: $M = 6$ and $M = 9$ players, $K = 9$ arms, horizon $T = 10000$ (for 1000 runs).

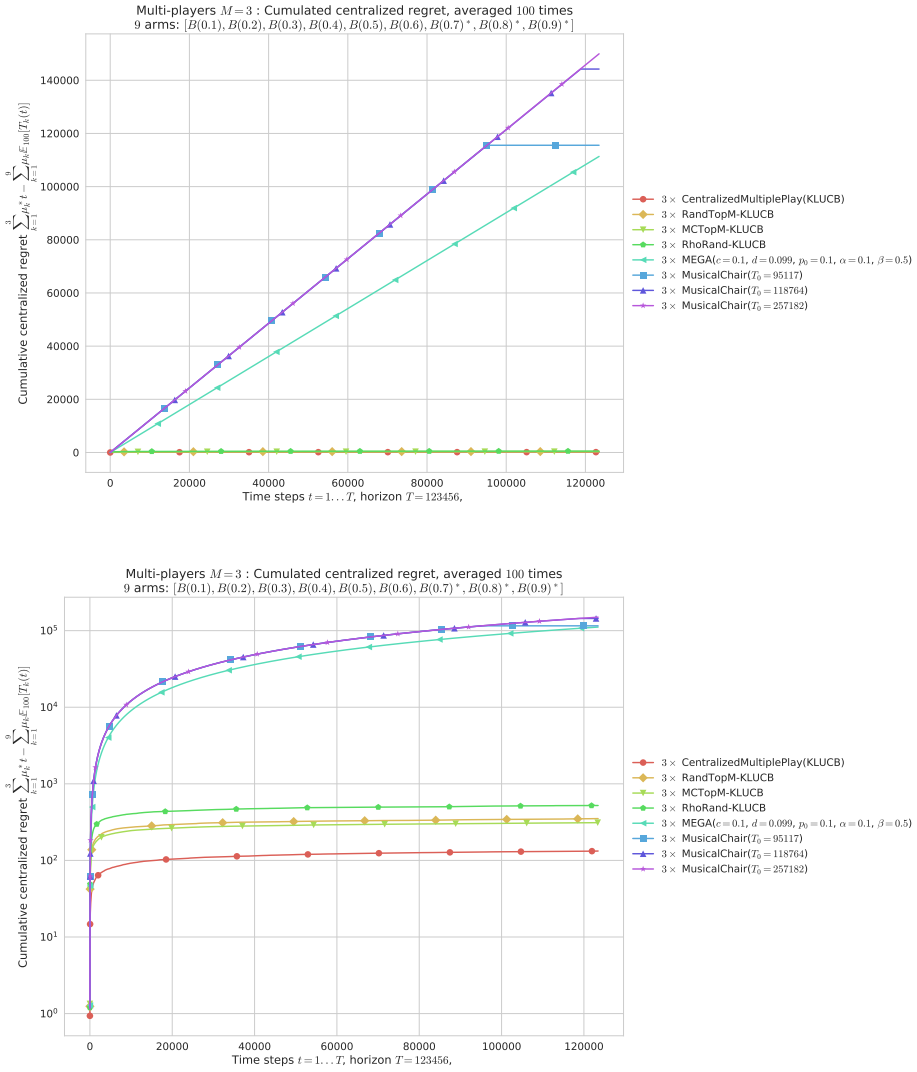


Figure 7: Regret (normal scale above, log-y below) for $M = 3$ players for $K = 9$ arms, horizon $T = 123456$, for 100 repetitions on problem $\mu = [0.1, \dots, 0.9]$. With a perfect knowledge on the gap ($\Delta = 0.1$ here) and by using the parameters suggested from their respective articles, MEGA and Musical Chair perform badly in this simple setting, even with the knowledge of the horizon T for Musical Chair. The first two Musical Chair instances use the optimal T_0 value from Rosenski et al. (2016), with ε taken slightly smaller than the gap Δ ($\varepsilon = 0.99\Delta$), and respectively with $\delta = 0.5$ and $\delta = 0.1$, for which the regret can bounded with probability 0.5 and 0.9 respectively. The third instance uses the optimal T_0 corresponding to $\delta = 1/T$, that is guaranteed to have an expected regret of order $\log(T)$.

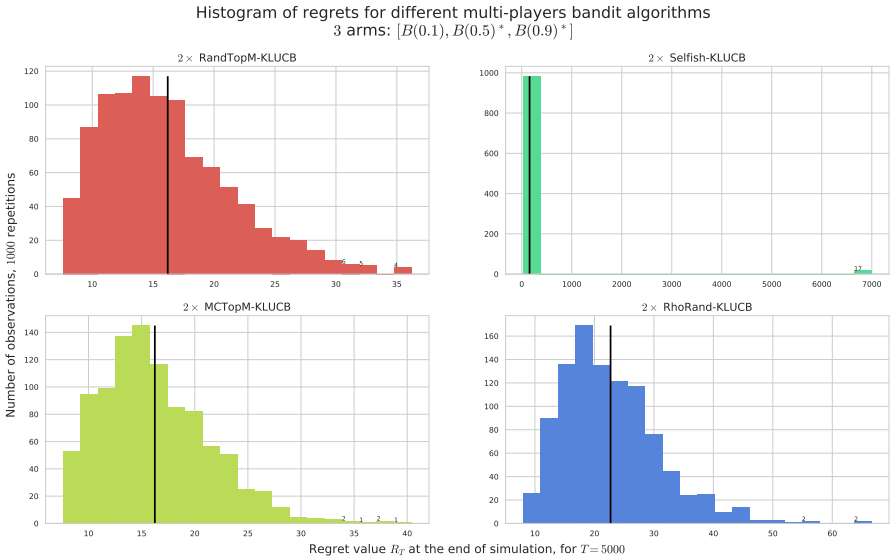


Figure 8: Regret for $M = 2$ players, $K = 3$ arms, horizon $T = 5000$, 1000 repetitions and $\boldsymbol{\mu} = [0.1, 0.5, 0.9]$. Axis x is for regret (different scale for each part), and the green curve for Selfish shows a small probability of having a linear regret (17 cases of $R_T \geq T$, out of 1000). The regret for the three other algorithms is very small for this problem, always smaller than 100 here.

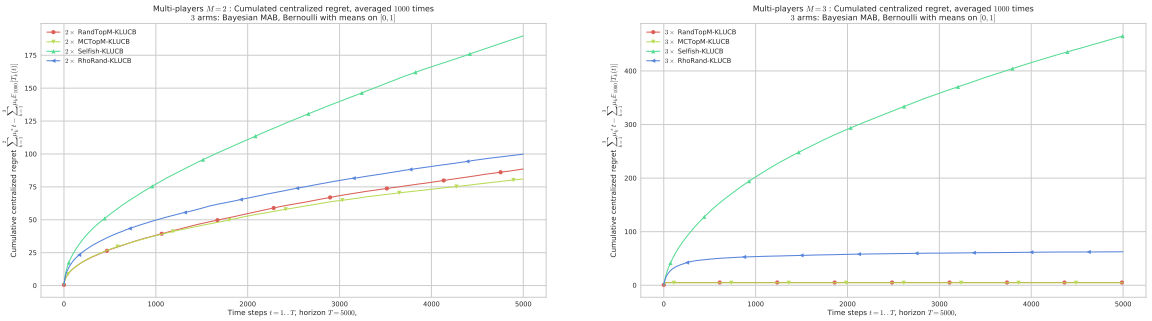


Figure 9: Regret, $M = 2$ and $M = 3$ players, $K = 3$ arms, horizon $T = 5000$, against 1000 problems $\boldsymbol{\mu}$ uniformly sampled in $[0, 1]^K$. Selfish (top curve in green) clearly fails in such setting with small K.

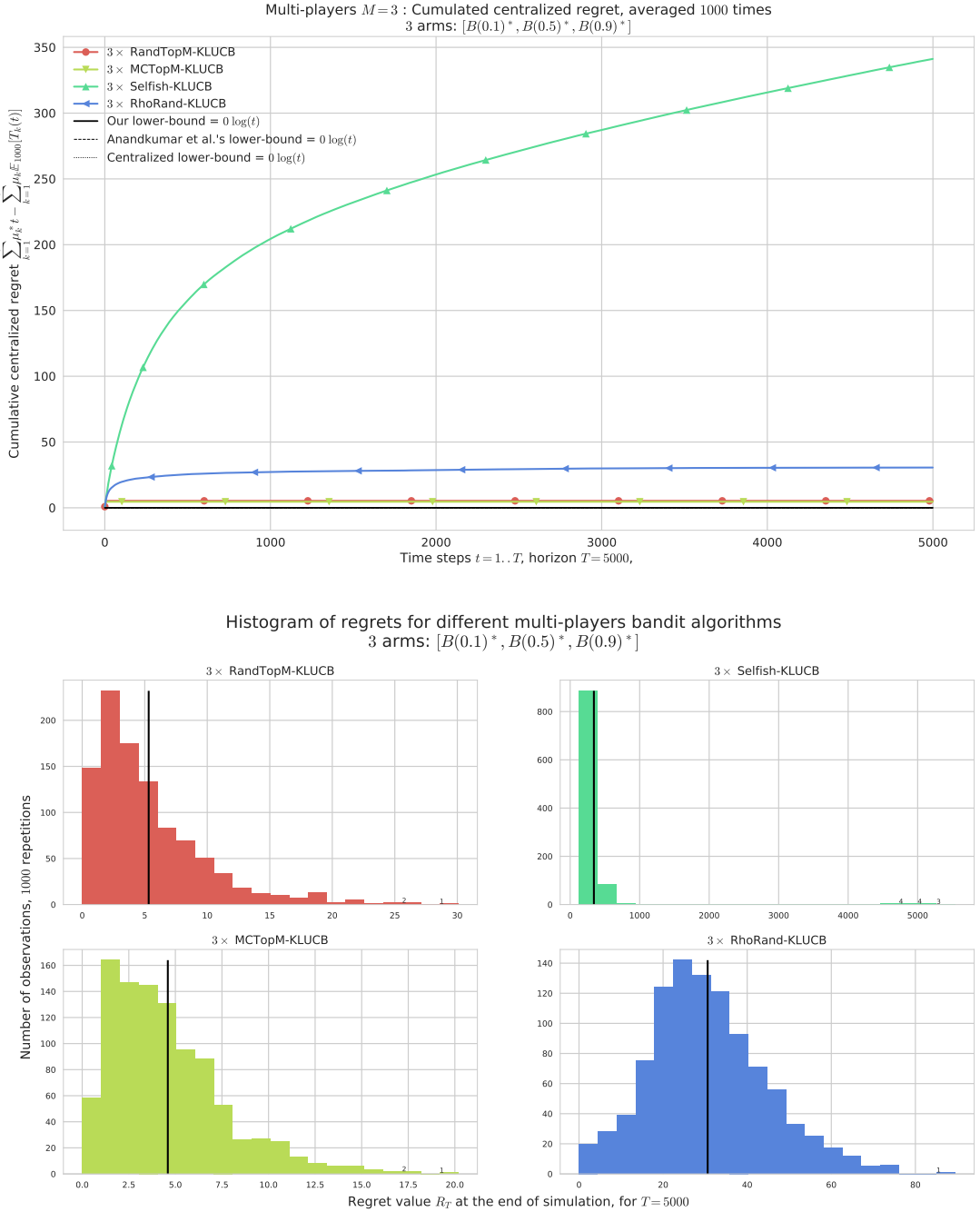


Figure 10: Regret for $M = 3$ players, $K = 3$ arms, horizon $T = 5000$, 1000 repetitions and $\boldsymbol{\mu} = [0.1, 0.5, 0.9]$. Axis x is for regret (different scale for each), and the top green curve for Selfish shows a small probability of having a linear regret (11 cases of $R_T \geq T$, out of 1000). The regret for the three other algorithms is very small for this problem, and even appears constant.

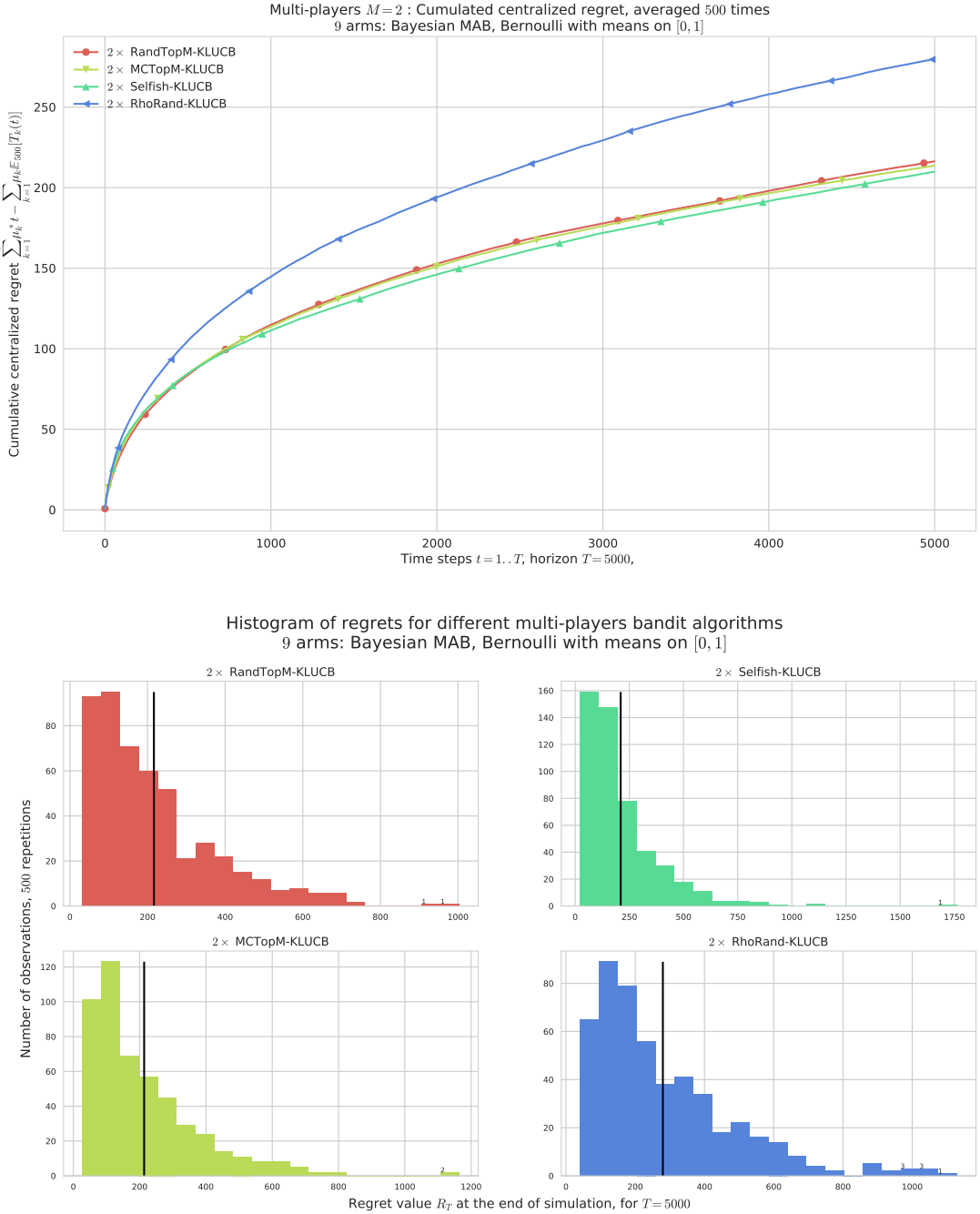


Figure 11: Regret, $M = 2$ players, $K = 9$ arms, horizon $T = 5000$, against 500 problems $\boldsymbol{\mu}$ uniformly sampled in $[0,1]^K$. RhoRand (top blue) is outperformed by the other algorithms (and the gain increases when M increases), which all perform similarly in such configurations. Note that the (small) tail of the histograms come from complicated problems $\boldsymbol{\mu}$ and not failure cases.

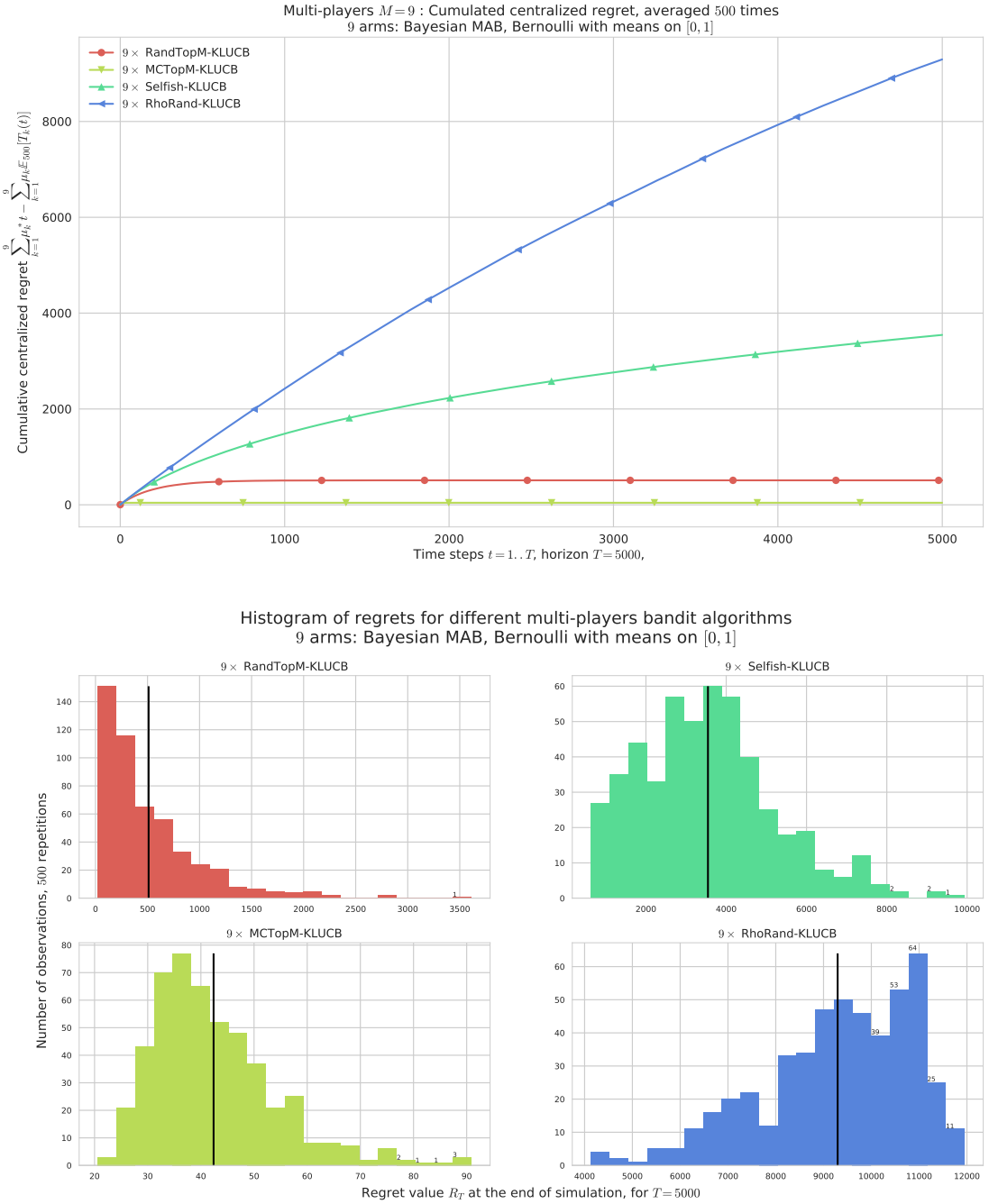


Figure 12: Regret, $M = 9$ players for $K = 9$ arms, horizon $T = 5000$, against 500 problems $\boldsymbol{\mu}$ uniformly sampled in $[0,1]^K$. This extreme case $M = K$ shows the drastic difference of behavior between RandTopM and MCTopM, having constant regret, and RhoRand and Selfish, having large regret.

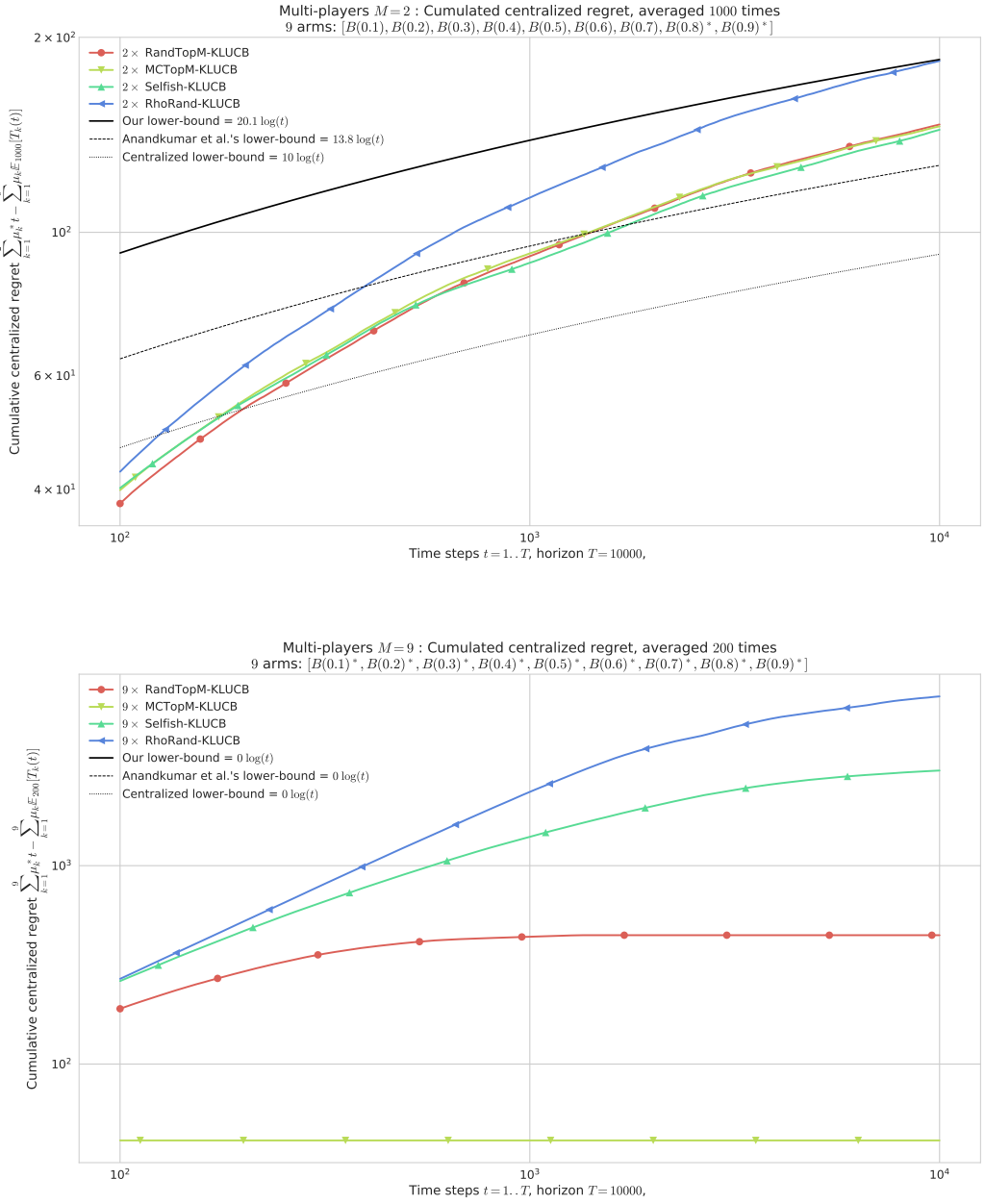


Figure 13: Regret (in $\log\log$ scale), for $M = 2$ and 9 players for $K = 9$ arms, horizon $T = 5000$, for problem $\boldsymbol{\mu} = [0.1, \dots, 0.9]$. In different settings, RandTopM (yellow curve) and Selfish (green) can outperform each other, and always outperform RhoRand. MCTopM is always among the best algorithms, and for M not too small, its regret seems logarithmic with a constant matching the lower bound.

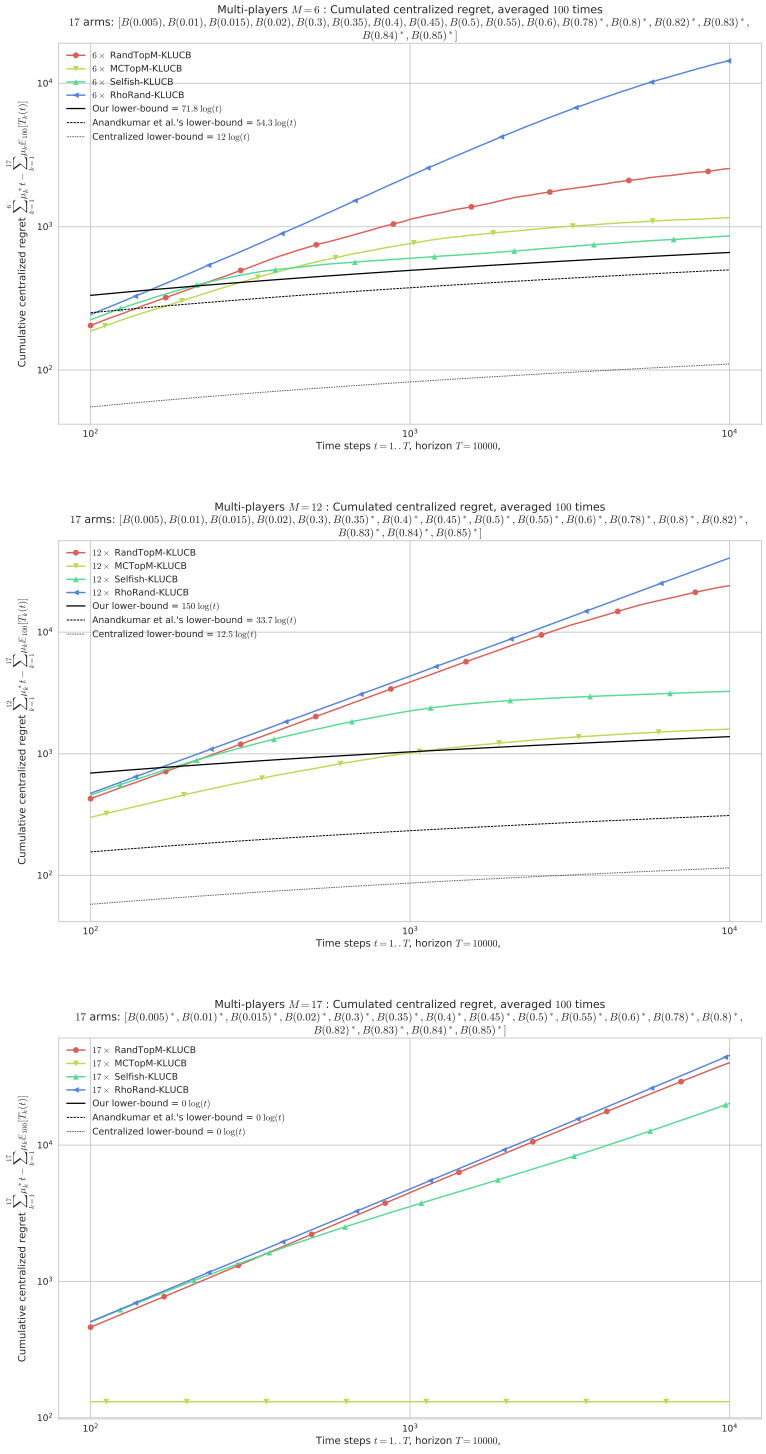


Figure 14: Regret (in log log scale), for $M = 6, 12, 17$ players for a "difficult" problem with $K = 17$, and $T = 5000$. The same observation as in Figure 13 can be made. Selfish outperforms MCTopM for $M = 2$ here. Additionally, MCTopM is the only algorithm to not fail dramatically when $M = K$ here.